

RIP-Based Near-Oracle Performance Guarantees for Subspace-Pursuit, CoSaMP, and Iterative Hard-Thresholding

Raja Giryes and Michael Elad

Department of Computer Science – Technion – Israel Institute of Technology

Haifa 32000, Israel

{raja,elad}@cs.technion.ac.il

Abstract

This paper presents an average case denoising performance analysis for the Subspace Pursuit (SP), the CoSaMP and the IHT algorithms. This analysis considers the recovery of a noisy signal, with the assumptions that (i) it is corrupted by an additive random white Gaussian noise; and (ii) it has a K -sparse representation with respect to a known dictionary \mathbf{D} . The proposed analysis is based on the Restricted-Isometry-Property (RIP), establishing a near-oracle performance guarantee for each of these algorithms. The results for the three algorithms differ in the bounds' constants and in the cardinality requirement (the upper bound on K for which the claim is true).

Similar RIP-based analysis was carried out previously for the Dantzig Selector (DS) and the Basis Pursuit (BP). Past work also considered a mutual-coherence-based analysis of the denoising performance of the DS, BP, the Orthogonal Matching Pursuit (OMP) and the thresholding algorithms. This work differs from the above as it addresses a different set of algorithms. Also, despite the fact that SP, CoSaMP, and IHT are greedy-like methods, the performance guarantees developed in this work resemble those obtained for the relaxation-based methods (DS and BP), suggesting that the performance is independent of the sparse representation entries contrast and magnitude.

I. INTRODUCTION

A. General – Pursuit Methods for Denoising

The area of sparse approximation (and compressed sensing as one prominent manifestation of its applicability) is an emerging field that get much attention in the last decade. In one of the most basic

problems posed in this field, we consider a noisy measurement vector $\mathbf{y} \in \mathbb{R}^m$ of the form

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e}, \quad (\text{I.1})$$

where $\mathbf{x} \in \mathbb{R}^N$ is the signal's representation with respect to the dictionary $\mathbf{D} \in \mathbb{R}^{m \times N}$ where $N \geq m$. The vector $\mathbf{e} \in \mathbb{R}^m$ is an additive noise, which is assumed to be an adversarial disturbance, or a random vector – e.g., white Gaussian noise with zero mean and variance σ^2 . We further assume that the columns of \mathbf{D} are normalized, and that the representation vector \mathbf{x} is K -sparse, or nearly so.¹ Our goal is to find the K -sparse vector \mathbf{x} that approximates the measured signal \mathbf{y} . Put formally, this reads

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \quad \text{subject to } \|\mathbf{x}\|_0 = K, \quad (\text{I.2})$$

where $\|\mathbf{x}\|_0$ is the ℓ_0 pseudo-norm that counts the number of non-zeros in the vector \mathbf{x} . This problem is quite hard and problematic [1], [2], [3], [4]. A straight forward search for the solution of (I.2) is an NP hard problem as it requires a combinatorial search over the support of \mathbf{x} [5]. For this reason, approximation algorithms were proposed – these are often referred to as pursuit algorithms.

One popular pursuit approach is based on ℓ_1 relaxation and known as the Basis Pursuit (BP) [6] or the Lasso [7]. The BP aims at minimizing the relaxed objective

$$(P1) : \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \leq \epsilon_{\text{BP}}^2, \quad (\text{I.3})$$

where ϵ_{BP} is a constant related to the noise power. This minimizing problem has an equivalent form:

$$(BP) : \min_{\mathbf{x}} \frac{1}{2} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 + \gamma_{BP} \|\mathbf{x}\|_1, \quad (\text{I.4})$$

where γ_{BP} is a constant related to ϵ_{BP} . Another ℓ_1 -based relaxed algorithm is the Dantzig Selector (DS), as proposed in [8]. The DS aims at minimizing

$$(DS) : \min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subject to } \|\mathbf{D}^*(\mathbf{y} - \mathbf{D}\mathbf{x})\|_\infty \leq \epsilon_{\text{DS}}, \quad (\text{I.5})$$

where ϵ_{DS} is a constant related to the noise power.

A different pursuit approach towards the approximation of the solution of (I.2) is the greedy strategy [9], [10], [11], leading to algorithms such as the Matching Pursuit (MP) and the Orthogonal Matching Pursuit (OMP). These algorithms build the solution \mathbf{x} one non-zero entry at a time, while greedily aiming to reduce the residual error $\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2$.

¹A more exact definition of nearly sparse vectors will be given later on

The last family of pursuit methods we mention here are greedy-like algorithms that differ from MP and OMP in two important ways: (i) Rather than accumulating the desired solution one element at a time, a group of non-zeros is identified together; and (ii) As opposed to the MP and OMP, these algorithms enable removal of elements from the detected support. Algorithms belonging to this group are the Regularized OMP (ROMP) [12], the Compressive Sampling Matching Pursuit (CoSaMP) [13], the Subspace-Pursuit (SP) [14], and the Iterative Hard Thresholding (IHT) [15]. This paper focuses on this specific family of methods, as it poses an interesting compromise between the simplicity of the greedy methods and the strong abilities of the relaxed algorithms.

B. Performance Analysis – Basic Tools

Recall that we aim at recovering the (deterministic!) sparse representation vector \mathbf{x} . We measure the quality of the approximate solution $\hat{\mathbf{x}}$ by the Mean-Squared-Error (MSE)

$$\text{MSE}(\hat{\mathbf{x}}) = E \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2, \quad (\text{I.6})$$

where the expectation is taken over the distribution of the noise. Therefore, our goal is to get as small as possible error. The question is, how small can this noise be? In order to answer this question, we first define two features that characterize the dictionary \mathbf{D} – the mutual coherence and the Restricted Isometry Property (RIP). Both are used extensively in formulating the performance guarantees of the sort developed in this paper.

The mutual-coherence μ [16], [17], [18] of a matrix \mathbf{D} is the largest absolute normalized inner product between different columns from \mathbf{D} . The larger it is, the more problematic the dictionary is, because in such a case we get that columns in \mathbf{D} are too much alike.

Turning to the RIP, it is said that \mathbf{D} satisfies the K -RIP condition with parameter δ_K if it is the smallest value that satisfies

$$(1 - \delta_K) \|\mathbf{x}\|_2^2 \leq \|\mathbf{D}\mathbf{x}\|_2^2 \leq (1 + \delta_K) \|\mathbf{x}\|_2^2 \quad (\text{I.7})$$

for any K -sparse vector \mathbf{x} [19], [20].

These two measures are related by $\delta_K \leq (K - 1)\mu$ [21]. The RIP is a stronger descriptor of \mathbf{D} as it characterizes groups of K columns from \mathbf{D} , whereas the mutual coherence “sees” only pairs. On the other hand, computing μ is easy, while the evaluation of δ_K is prohibitive in most cases. An exception to this are random matrices \mathbf{D} for which the RIP constant is known (with high probability). For example,

if the entries of $\sqrt{m}\mathbf{D}$ are drawn from a white Gaussian distribution² and $m \geq CK \log(N/K)/\epsilon^2$, then with a very high probability $\delta_K \leq \epsilon$ [19], [22].

We return now to the question we posed above: how small can the error $\text{MSE}(\hat{\mathbf{x}})$ be? Consider an oracle estimator that knows the support of \mathbf{x} , i.e. the locations of the K non-zeros in this vector. The oracle estimator obtained as a direct solution of the problem posed in (I.2) is easily given by

$$\hat{\mathbf{x}}_{\text{oracle}} = \mathbf{D}_T^\dagger \mathbf{y}, \quad (\text{I.8})$$

where T is the support of \mathbf{x} and \mathbf{D}_T is a sub-matrix of \mathbf{D} that contains only the columns involved in the support T . Its MSE is given by [8]

$$\text{MSE}(\hat{\mathbf{x}}_{\text{oracle}}) = E \|\mathbf{x} - \hat{\mathbf{x}}_{\text{oracle}}\|_2^2 = E \|\mathbf{D}_T^\dagger \mathbf{e}\|_2^2. \quad (\text{I.9})$$

In the case of a random noise, as described above, this error becomes

$$\begin{aligned} \text{MSE}(\hat{\mathbf{x}}_{\text{oracle}}) &= E \|\mathbf{D}_T^\dagger \mathbf{e}\|_2^2 \\ &= \text{trace} \{ (\mathbf{D}_T^* \mathbf{D}_T)^{-1} \} \sigma^2 \\ &\leq \frac{K}{1 - \delta_K} \sigma^2. \end{aligned} \quad (\text{I.10})$$

This is the smallest possible error, and it is proportional to the number of non-zeros K multiplied by σ^2 . It is natural to ask how close do we get to this best error by practical pursuit methods that do not assume the knowledge of the support. This brings us to the next sub-section.

C. Performance Analysis – Known Results

There are various attempts to bound the MSE of pursuit algorithms. Early works considered the adversary case, where the noise can admit any form as long as its norm is bounded [23], [24], [2], [1]. These works gave bounds on the reconstruction error in the form of a constant factor ($\text{Const} > 1$) multiplying the noise power,

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq \text{Const} \cdot \|\mathbf{e}\|_2^2. \quad (\text{I.11})$$

Notice that the cardinality of the representation plays no role in this bound, and all the noise energy is manifested in the final error.

²The multiplication by \sqrt{m} comes to normalize the columns of the effective dictionary \mathbf{D} .

One such example is the work by Candès and Tao, reported in [20], which analyzed the BP error. This work have shown that if the dictionary \mathbf{D} satisfies $\delta_K + \delta_{2K} + \delta_{3K} < 1$ then the BP MSE is bounded by a constant times the energy of the noise, as shown above. The condition on the RIP was improved to $\delta_{2K} < \sqrt{2} - 1$ in [25]. Similar tighter bounds are $\delta_{1.625K} < \sqrt{2} - 1$ and $\delta_{3K} < 4 - 2\sqrt{3}$ [26], or $\delta_K < 0.307$ [27]. The advantage of using the RIP in the way described above is that it gives a uniform guarantee: it is related only to the dictionary and sparsity level.

Next in line to be analyzed are the greedy methods (MP, OMP, Thr) [23], [1]. Unlike the BP, these algorithms where shown to be more sensitive, incapable of providing a uniform guarantee for the reconstruction. Rather, beyond the dependence on the properties of \mathbf{D} and the sparsity level, the guarantees obtained depend also on the ratio between the noise power and the absolute values of the signal representation entries.

Interestingly, the greedy-like approach, as practiced in the ROMP, the CoSaMP, the SP, and the IHT algorithms, was found to be closer in spirit to the BP, all leading to uniform guarantees on the bounded MSE. The ROMP was the first of these algorithms to be analyzed [12], leading to the more strict requirement $\delta_{2K} < 0.03/\sqrt{\log K}$. The CoSaMP [13] and the SP [14] that came later have similar RIP conditions without the $\log K$ factor, where the SP result is slightly better. The IHT algorithm was also shown to have a uniform guarantee for bounded error of the same flavor as shown above [15].

All the results mentioned above deal with an adversarial noise, and therefore give bounds that are related only to the noise power with a coefficient that is larger than 1, implying that no effective denoising is to be expected. This is natural since we consider the worst case results, where the noise can be concentrated in the places of the non-zero elements of the sparse vector. To obtain better results, one must change the perspective and consider a random noise drawn from a certain distribution.

The first to realize this and exploit this alternative point of view were Candès and Tao in the work reported in [8] that analyzed the DS algorithm. As mentioned above, the noise was assumed to be random zero-mean white Gaussian noise with a known variance σ^2 . For the choice $\epsilon_{DS} = \sqrt{2(1+a)\log N} \cdot \sigma$, and requiring $\delta_{2K} + \delta_{3K} < 1$, the minimizer of (I.5), $\hat{\mathbf{x}}_{DS}$, was shown to obey

$$\|\mathbf{x} - \hat{\mathbf{x}}_{DS}\|_2^2 \leq C_{DS}^2 \cdot (2(1+a)\log N) \cdot K\sigma^2, \quad (\text{I.12})$$

with probability exceeding $1 - (\sqrt{\pi(1+a)\log N} \cdot N^a)^{-1}$, where $C_{DS} = 4/(1 - 2\delta_{3K})$.³ Up to a constant and a $\log N$ factor, this bound is the same as the oracle's one in (I.9). The $\log N$ factor in (I.12)

³In [8] a slightly different constant was presented.

in unavoidable, as proven in [28], and therefore this bound is optimal up to a constant factor.

A similar result was presented in [29] for the BP, showing that the solution of (I.4) for the choice $\gamma_{BP} = \sqrt{8\sigma^2(1+a)\log N}$, and requiring $\delta_{2K} + 3\delta_{3K} < 1$, satisfies

$$\|\mathbf{x} - \hat{\mathbf{x}}_{DS}\|_2^2 \leq C_{BP}^2 \cdot (2(1+a)\log N) \cdot K\sigma^2 \quad (\text{I.13})$$

with probability exceeding $1 - (N^a)^{-1}$. This result is weaker than the one obtained for the DS in three ways: (i) It gives a smaller probability of success; (ii) The constant C_{BP} is larger, as shown in [21] ($C_{BP} \geq 32/\kappa^4$, where $\kappa < 1$ is defined in [29]); and (iii) The condition on the RIP is stronger.

Mutual-Coherence based results for the DS and BP were derived in [30], [21]. In [21] results were developed also for greedy algorithms – the OMP and the thresholding. These results rely on the contrast and magnitude of the entries of \mathbf{x} . Denoting by $\hat{\mathbf{x}}_{greedy}$ the reconstruction result of the thresholding and the OMP, we have

$$\|\mathbf{x} - \hat{\mathbf{x}}_{greedy}\|_2^2 \leq C_{greedy}^2 \cdot (2(1+a)\log N) \cdot K\sigma^2, \quad (\text{I.14})$$

where $C_{greedy} \leq 2$ and with probability exceeds $1 - (\sqrt{\pi(1+a)\log N} \cdot N^a)^{-1}$. This result is true for the OMP and thresholding under the condition

$$\frac{|\mathbf{x}_{min}| - 2\sigma\sqrt{2(1+a)\log N}}{(2K-1)\mu} \geq \begin{cases} |\mathbf{x}_{min}| & \text{OMP} \\ |\mathbf{x}_{max}| & \text{THR} \end{cases}, \quad (\text{I.15})$$

where $|\mathbf{x}_{min}|$ and $|\mathbf{x}_{max}|$ are the minimal and maximal non-zero absolute entries in \mathbf{x} .

D. This Paper Contribution

We have seen that greedy algorithms' success is dependent on the magnitude of the entries of \mathbf{x} and the noise power, which is not the case for the DS and BP. It seems that there is a need for pursuit algorithms that, on one hand, will enjoy the simplicity and ease of implementation as in the greedy methods, while being guaranteed to perform as well as the BP and DS. Could the greedy-like methods (ROMP, CoSaMP, SP, IHT) serve this purpose? The answer was shown to be positive for the adversarial noise assumption, but these results are too weak, as they do not show the true denoising effect that such algorithm may lead to. In this work we show that the answer remains positive for the random noise assumption.

More specifically, in this paper we present RIP-based near-oracle performance guarantees for the SP, CoSaMP and IHT algorithms (in this order). We show that these algorithms get uniform guarantees, just as for the relaxation based methods (the DS and BP). We present the analysis that leads to these results

and we provide explicit values for the constants in the obtained bounds.

The organization of this paper is as follows: In Section II we introduce the notation and propositions used for our analysis. In Section III we develop RIP-based bounds for the SP, CoSaMP and the IHT algorithms for the adversarial case. Then we show how we can derive from these a new set of guarantees for near oracle performance that consider the noise as random. We develop fully the steps for the SP, and outline the steps needed to get the results for the CoSaMP and IHT. In Section IV we present some experiments that show the performance of the three methods, and a comparison between the theoretical bounds and the empirical performance. In Section V we consider the nearly-sparse case, extending all the above results. Section VI concludes our work.

II. NOTATION AND PRELIMINARIES

The following notations are used in this paper:

- $\text{supp}(\mathbf{x})$ is the support of \mathbf{x} (a set with the locations of the non-zero elements of \mathbf{x}).
- $|\text{supp}(\mathbf{x})|$ is the size of the set $\text{supp}(\mathbf{x})$.
- $\text{supp}(\mathbf{x}, K)$ is the support of the largest K magnitude elements in \mathbf{x} .
- \mathbf{D}_T is a matrix composed of the columns of the matrix \mathbf{D} of the set T .
- In a similar way, \mathbf{x}_T is a vector composed of the entries of the vector \mathbf{x} over the set T .
- T^C symbolizes the complementary set of T .
- $T - \tilde{T}$ is the set of all the elements contained in T but not in \tilde{T} .
- We will denote by T the set of the non-zero places of the original signal \mathbf{x} ; As such, $|T| \leq K$ when \mathbf{x} is K -sparse.
- \mathbf{x}_K is the vector with the K dominant elements of \mathbf{x} .
- The projection of a vector \mathbf{y} to the subspace spanned by the columns of the matrix \mathbf{A} (assumed to have more rows than columns) is denoted by $\text{proj}(\mathbf{y}, \mathbf{A}) = \mathbf{A}\mathbf{A}^\dagger \mathbf{y}$. The residual is $\text{resid}(\mathbf{y}, \mathbf{A}) = \mathbf{y} - \mathbf{A}\mathbf{A}^\dagger \mathbf{y}$.
- T_e is the subset of columns of size K in \mathbf{D} that gives the maximum correlation with the noise vector \mathbf{e} , namely,

$$T_e = \underset{T \mid |T|=K}{\text{argmax}} \|\mathbf{D}_T^* \mathbf{e}\|_2 \quad (\text{II.1})$$

- $T_{e,p}$ is a generalization of T_e where T in (II.1) is of size pK , $p \in \mathbb{N}$. It is clear that $\|\mathbf{D}_{T_{e,p}}^* \mathbf{e}\|_2 \leq p \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2$.

The proofs in this paper use several propositions from [13], [14]. We bring these in this Section, so as to keep the discussion complete.

Proposition 2.1: [Proposition 3.1 in [13]] Suppose \mathbf{D} has a restricted isometry constant δ_K . Let T be a set of K indices or fewer. Then

$$\begin{aligned}\|\mathbf{D}_T^* \mathbf{y}\|_2 &\leq \sqrt{1 + \delta_K} \|\mathbf{y}\|_2 \\ \|\mathbf{D}_T^\dagger \mathbf{y}\|_2 &\leq \frac{1}{\sqrt{1 - \delta_K}} \|\mathbf{y}\|_2 \\ \|\mathbf{D}_T^* \mathbf{D}_T \mathbf{x}\|_2 &\stackrel{\leq}{\geq} (1 \pm \delta_K) \|\mathbf{x}\|_2 \\ \|(\mathbf{D}_T^* \mathbf{D}_T)^{-1} \mathbf{x}\|_2 &\stackrel{\leq}{\geq} \frac{1}{1 \pm \delta_K} \|\mathbf{x}\|_2\end{aligned}$$

where the last two statements contain upper and lower bounds, depending on the sign chosen.

Proposition 2.2: [Lemma 1 in [14]] Consequences of the RIP:

- 1) (Monotonicity of δ_K) For any two integers $K \leq K'$, $\delta_K \leq \delta_{K'}$.
- 2) (Near-orthogonality of columns) Let $I, J \subset \{1, \dots, N\}$ be two disjoint sets ($I \cap J = \emptyset$). Suppose that $\delta_{|I|+|J|} < 1$. For arbitrary vectors $\mathbf{a} \in \mathbb{R}^{|I|}$ and $\mathbf{b} \in \mathbb{R}^{|J|}$,

$$|\langle \mathbf{D}_I \mathbf{a}, \mathbf{D}_J \mathbf{b} \rangle| \leq \delta_{|I|+|J|} \|\mathbf{a}\|_2 \|\mathbf{b}\|_2$$

and

$$\|\mathbf{D}_I^* \mathbf{D}_J \mathbf{b}\|_2 \leq \delta_{|I|+|J|} \|\mathbf{b}\|_2.$$

Proposition 2.3: [Lemma 2 in [14]] Projection and Residue:

- 1) (Orthogonality of the residue) For an arbitrary vector $\mathbf{y} \in \mathbb{R}^m$ and a sub-matrix $\mathbf{D}_I \in \mathbb{R}^{m \times K}$ of full column-rank, let $\mathbf{y}_r = \text{resid}(\mathbf{y}, \mathbf{D}_I)$. Then $\mathbf{D}_I^* \mathbf{y}_r = 0$.
- 2) (Approximation of the projection residue) Consider a matrix $\mathbf{D} \in \mathbb{R}^{m \times N}$. Let $I, J \subset \{1, \dots, N\}$ be two disjoint sets, $I \cap J = \emptyset$, and suppose that $\delta_{|I|+|J|} < 1$. Let $\mathbf{y} \in \text{span}(\mathbf{D}_I)$, $\mathbf{y}_p = \text{proj}(\mathbf{y}, \mathbf{D}_J)$ and $\mathbf{y}_r = \text{resid}(\mathbf{y}, \mathbf{D}_J)$. Then

$$\|\mathbf{y}_p\|_2 \leq \frac{\delta_{|I|+|J|}}{1 - \delta_{\max(|I|, |J|)}} \|\mathbf{y}\|_2$$

and

$$\left(1 - \frac{\delta_{|I|+|J|}}{1 - \delta_{\max(|I|, |J|)}}\right) \|\mathbf{y}\|_2 \leq \|\mathbf{y}_r\|_2 \leq \|\mathbf{y}\|_2.$$

Proposition 2.4: [Corollary 3.3 in [13]] Suppose that \mathbf{D} has an RIP constant $\delta_{\tilde{K}}$. Let T_1 be an arbitrary set of indices, and let \mathbf{x} be a vector. Provided that $\tilde{K} \geq |T_1 \cup \text{supp}(\mathbf{x})|$, we obtain that

$$\|\mathbf{D}_{T_1}^* \mathbf{D}_{T_1^c} \mathbf{x}_{T_1^c}\|_2 \leq \delta_{\tilde{K}} \|\mathbf{x}_{T_1^c}\|_2. \quad (\text{II.2})$$

III. NEAR ORACLE PERFORMANCE OF THE ALGORITHMS

Our goal in this section is to find error bounds for the SP, CoSaMP and IHT reconstructions given the measurement from (I.1). We will first find bounds for the case where \mathbf{e} is an adversarial noise using the same techniques used in [14], [13]. In these works and in [15], the reconstruction error was bounded by a constant times the noise power in the same form as in (I.11). In this work, we will derive a bound that is a constant times $\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2$ (where T_e is as defined in the previous section). Armed with this bound, we will change perspective and look at the case where \mathbf{e} is a white Gaussian noise, and derive a near-oracle performance result of the same form as in (I.12), using the same tools used in [8].

A. Near oracle performance of the SP algorithm

We begin with the SP pursuit method, as described in Algorithm 1. SP holds a temporal solution with K non-zero entries, and in each iteration it adds an additional set of K candidate non-zeros that are most correlated with the residual, and prunes this list back to K elements by choosing the dominant ones. We use a constant number of iterations as a stopping criterion but different stopping criteria can be sought, as presented in [14].

Algorithm 1 Subspace Pursuit Algorithm [Algorithm 1 in [14]]

Input: $K, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e}$, K is the cardinality of \mathbf{x} and \mathbf{e} is the additive noise.

Output: $\hat{\mathbf{x}}_{SP}$: K -sparse approximation of \mathbf{x}

Initialize the support: $T^0 = \emptyset$.

Initialize the residual: $\mathbf{y}_r^0 = \mathbf{y}$.

while halting criterion is not satisfied **do**

Find new support elements: $T_\Delta = \text{supp}(\mathbf{D}^* \mathbf{y}_r^{\ell-1}, K)$.

Update the support: $\tilde{T}^\ell = T^{\ell-1} \cup T_\Delta$.

Compute the representation: $\mathbf{x}_p = \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{y}$.

Prune small entries in the representation: $T^\ell = \text{supp}(\mathbf{x}_p, K)$.

Update the residual: $\mathbf{y}_r^\ell = \text{resid}(\mathbf{y}, \mathbf{D}_{T^\ell})$.

end while

Form the final solution: $\hat{\mathbf{x}}_{SP, (T^\ell)^c} = 0$ and $\hat{\mathbf{x}}_{SP, T^\ell} = \mathbf{D}_{T^\ell}^\dagger \mathbf{y}$.

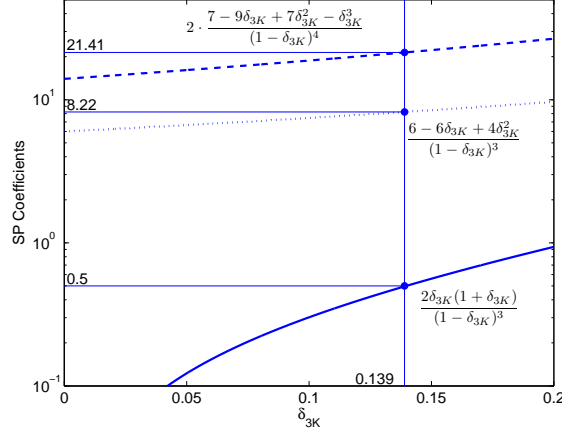


Fig. 1. The coefficients in (III.1) and (III.5) as functions of δ_{3K} .

Theorem 3.1: The SP solution at the ℓ -th iteration satisfies the recurrence inequality

$$\begin{aligned} \|\mathbf{x}_{T-T^\ell}\|_2 &\leq \frac{2\delta_{3K}(1+\delta_{3K})}{(1-\delta_{3K})^3} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 \\ &\quad + \frac{6-6\delta_{3K}+4\delta_{3K}^2}{(1-\delta_{3K})^3} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{III.1})$$

For $\delta_{3K} \leq 0.139$ this leads to

$$\|\mathbf{x}_{T-T^\ell}\|_2 \leq 0.5 \|\mathbf{x}_{T-T^{\ell-1}}\|_2 + 8.22 \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \quad (\text{III.2})$$

Proof: The proof of the inequality in (III.1) is given in Appendix A. Note that the recursive formula given (III.1) has two coefficients, both functions of δ_{3K} . Fig. 1 shows these coefficients as a function of δ_{3K} . As can be seen, under the condition $\delta_{3K} \leq 0.139$, it holds that the coefficient multiplying $\|\mathbf{x}_{T-T^{\ell-1}}\|_2$ is lesser or equal to 0.5, while the coefficient multiplying $\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2$ is lesser or equal to 8.22, which completes our proof. \square

Corollary 3.2: Under the condition $\delta_{3K} \leq 0.139$, the SP algorithm satisfies

$$\|\mathbf{x}_{T-T^\ell}\|_2 \leq 2^{-\ell} \|\mathbf{x}\|_2 + 2 \cdot 8.22 \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \quad (\text{III.3})$$

In addition, After at most

$$\ell^* = \left\lceil \log_2 \left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2} \right) \right\rceil \quad (\text{III.4})$$

iterations, the solution $\hat{\mathbf{x}}_{SP}$ leads to an accuracy

$$\|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2 \leq C_{SP} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2, \quad (\text{III.5})$$

where

$$C_{SP} = 2 \cdot \frac{7 - 9\delta_{3K} + 7\delta_{3K}^2 - \delta_{3K}^3}{(1 - \delta_{3K})^4} \leq 21.41 \quad (\text{III.6})$$

Proof: Starting with (III.2), and applying it recursively we obtain

$$\begin{aligned} \|\mathbf{x}_{T-T^\ell}\|_2 &\leq 0.5 \|\mathbf{x}_{T-T^{\ell-1}}\|_2 + 8.22 \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\ &\leq 0.5^2 \|\mathbf{x}_{T-T^{\ell-2}}\|_2 + 8.22 \cdot (0.5 + 1) \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\ &\leq \dots \\ &\leq 0.5^k \|\mathbf{x}_{T-T^{\ell-k}}\|_2 + 8.22 \cdot \left(\sum_{j=0}^{k-1} 0.5^j \right) \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{III.7})$$

Setting $k = \ell$ leads easily to (III.3), since $\|\mathbf{x}_{T-T^0}\|_2 = \|\mathbf{x}_T\|_2 = \|\mathbf{x}\|_2$.

Plugging the number of iterations ℓ^* as in (III.4) to (III.3) yields⁴

$$\begin{aligned} \|\mathbf{x}_{T-T^{\ell^*}}\|_2 & \\ &\leq 2^{-\ell^*} \|\mathbf{x}\|_2 + 2 \cdot \frac{6 - 6\delta_{3K} + 4\delta_{3K}^2}{(1 - \delta_{3K})^3} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\ &\leq \left(1 + 2 \cdot \frac{6 - 6\delta_{3K} + 4\delta_{3K}^2}{(1 - \delta_{3K})^3} \right) \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{III.8})$$

We define $\hat{T} \triangleq T^{\ell^*}$ and bound the reconstruction error $\|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2$. First, notice that $\|\mathbf{x}\| = \|\mathbf{x}_{\hat{T}}\| + \|\mathbf{x}_{T-\hat{T}}\|$, simply because the true support T can be divided into⁵ \hat{T} and the complementary part, $T - \hat{T}$.

Using the facts that $\hat{\mathbf{x}}_{SP} = \mathbf{D}_{\hat{T}}^\dagger \mathbf{y}$, $\mathbf{y} = \mathbf{D}_T \mathbf{x}_T + \mathbf{e}$, and the triangle inequality, we get

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2 & \\ &\leq \left\| \mathbf{x}_{\hat{T}} - \mathbf{D}_{\hat{T}}^\dagger \mathbf{y} \right\|_2 + \left\| \mathbf{x}_{T-\hat{T}} \right\|_2 \\ &= \left\| \mathbf{x}_{\hat{T}} - \mathbf{D}_{\hat{T}}^\dagger (\mathbf{D}_T \mathbf{x}_T + \mathbf{e}) \right\|_2 + \left\| \mathbf{x}_{T-\hat{T}} \right\|_2 \\ &\leq \left\| \mathbf{x}_{\hat{T}} - \mathbf{D}_{\hat{T}}^\dagger \mathbf{D}_T \mathbf{x}_T \right\|_2 + \left\| \mathbf{D}_{\hat{T}}^\dagger \mathbf{e} \right\|_2 + \left\| \mathbf{x}_{T-\hat{T}} \right\|_2. \end{aligned} \quad (\text{III.9})$$

⁴Note that we have replaced the constant 8.22 with the equivalent expression that depends on δ_{3K} – see (III.1).

⁵The vector $\mathbf{x}_{\hat{T}}$ is of length $|\hat{T}| = K$ and it contains zeros in locations that are outside T .

We proceed by breaking the term $\mathbf{D}_T \mathbf{x}_T$ into the sum $\mathbf{D}_{T \cap \hat{T}} \mathbf{x}_{T \cap \hat{T}} + \mathbf{D}_{T - \hat{T}} \mathbf{x}_{T - \hat{T}}$, and obtain

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2 &\leq \left\| \mathbf{x}_{\hat{T}} - \mathbf{D}_{\hat{T}}^\dagger \mathbf{D}_{T \cap \hat{T}} \mathbf{x}_{T \cap \hat{T}} \right\|_2 \\ &\quad + \left\| \mathbf{D}_{\hat{T}}^\dagger \mathbf{D}_{T - \hat{T}} \mathbf{x}_{T - \hat{T}} \right\|_2 + \left\| (\mathbf{D}_{\hat{T}}^* \mathbf{D}_{\hat{T}})^{-1} \mathbf{D}_{\hat{T}}^* \mathbf{e} \right\|_2 + \left\| \mathbf{x}_{T - \hat{T}} \right\|_2. \end{aligned} \quad (\text{III.10})$$

The first term in the above inequality vanishes, since $\mathbf{D}_{T \cap \hat{T}} \mathbf{x}_{T \cap \hat{T}} = \mathbf{D}_{\hat{T}} \mathbf{x}_{\hat{T}}$ (recall that $\mathbf{x}_{\hat{T}}$ outside the support T has zero entries that do not contribute to the multiplication). Thus, we get that $\mathbf{x}_{\hat{T}} - \mathbf{D}_{\hat{T}}^\dagger \mathbf{D}_{T \cap \hat{T}} \mathbf{x}_{T \cap \hat{T}} = \mathbf{x}_{\hat{T}} - \mathbf{D}_{\hat{T}}^\dagger \mathbf{D}_{\hat{T}} \mathbf{x}_{\hat{T}} = 0$. The second term can be bounded using Propositions 2.1 and 2.2,

$$\begin{aligned} \left\| \mathbf{D}_{\hat{T}}^\dagger \mathbf{D}_{T - \hat{T}} \mathbf{x}_{T - \hat{T}} \right\|_2 &= \left\| (\mathbf{D}_{\hat{T}}^* \mathbf{D}_{\hat{T}})^{-1} \mathbf{D}_{\hat{T}}^* \mathbf{D}_{T - \hat{T}} \mathbf{x}_{T - \hat{T}} \right\|_2 \\ &\leq \frac{1}{1 - \delta_K} \left\| \mathbf{D}_{\hat{T}}^* \mathbf{D}_{T - \hat{T}} \mathbf{x}_{T - \hat{T}} \right\|_2 \leq \frac{\delta_{2K}}{1 - \delta_K} \left\| \mathbf{x}_{T - \hat{T}} \right\|_2. \end{aligned}$$

Similarly, the third term is bounded using Propositions 2.1, and we obtain

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2 &\leq \left(1 + \frac{\delta_{2K}}{1 - \delta_K} \right) \left\| \mathbf{x}_{T - \hat{T}} \right\|_2 + \frac{1}{1 - \delta_K} \left\| \mathbf{D}_{\hat{T}}^* \mathbf{e} \right\|_2 \\ &\leq \frac{1}{1 - \delta_{3K}} \left\| \mathbf{x}_{T - \hat{T}} \right\|_2 + \frac{1}{1 - \delta_{3K}} \left\| \mathbf{D}_{\hat{T}}^* \mathbf{e} \right\|_2, \end{aligned}$$

where we have replaced δ_K and δ_{2K} with δ_{3K} , thereby bounding the existing expression from above.

Plugging (III.8) into this inequality leads to

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2 &\leq \frac{1}{1 - \delta_{3K}} \left(2 + 2 \cdot \frac{6 - 6\delta_{3K} + 4\delta_{3K}^2}{(1 - \delta_{3K})^3} \right) \left\| \mathbf{D}_{\hat{T}}^* \mathbf{e} \right\|_2 \\ &= 2 \cdot \frac{7 - 9\delta_{3K} + 7\delta_{3K}^2 - \delta_{3K}^3}{(1 - \delta_{3K})^4} \left\| \mathbf{D}_{\hat{T}}^* \mathbf{e} \right\|_2. \end{aligned}$$

Applying the condition $\delta_{3K} \leq 0.139$ on this equation leads to the result in (III.5). \square

For practical use we may suggest a simpler term for ℓ^* . Since $\left\| \mathbf{D}_{T_a}^* \mathbf{e} \right\|_2$ is defined by the subset that gives the maximal correlation with the noise, and it appears in the denominator of ℓ^* , it can be replaced with the average correlation, thus $\ell^* \approx \left\lceil \log_2 \left(\|\mathbf{x}\|_2 / \sqrt{K} \sigma \right) \right\rceil$.

Now that we have a bound for the SP algorithm for the adversarial case, we proceed and consider a bound for the random noise case, which will lead to a near-oracle performance guarantee for the SP algorithm.

Theorem 3.3: Assume that \mathbf{e} is a white Gaussian noise vector with variance σ^2 and that the columns of \mathbf{D} are normalized. If the condition $\delta_{3K} \leq 0.139$ holds, then with probability exceeding $1 - (\sqrt{\pi(1+a)} \log N)$.

$N^a)^{-1}$ we obtain

$$\|\mathbf{x} - \hat{\mathbf{x}}_{SP}\|_2^2 \leq C_{SP}^2 \cdot (2(1+a) \log N) \cdot K \sigma^2. \quad (\text{III.11})$$

Proof: Following Section 3 in [8] it holds true that $\mathbf{P} \left(\sup_i |\mathbf{D}_i^* \mathbf{e}| > \sigma \cdot \sqrt{2(1+a) \log N} \right) \leq 1 - (\sqrt{\pi(1+a) \log N} \cdot N^a)^{-1}$. Combining this with (III.5), and bearing in mind that $|T_e| = K$, we get the stated result. \square

As can be seen, this result is similar to the one posed in [8] for the Dantzig-Selector, but with a different constant – the one corresponding to DS is ≈ 5.5 for the RIP requirement used for the SP. For both algorithms, smaller values of δ_{3K} provide smaller constants.

B. Near oracle performance of the CoSaMP algorithm

We continue with the CoSaMP pursuit method, as described in Algorithm 2. CoSaMP, in a similar way to the SP, holds a temporal solution with K non-zero entries, with the difference that in each iteration it adds an additional set of $2K$ (instead of K) candidate non-zeros that are most correlated with the residual. Another difference is that after the pruning step in SP we use a matrix inversion in order to calculate a new projection for the K dominant elements, while in the CoSaMP we just take the biggest K elements. Here also, we use a constant number of iterations as a stopping criterion while different stopping criteria can be sought, as presented in [13].

Algorithm 2 CoSaMP Algorithm [Algorithm 2.1 in [13]]

Input: $K, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e}$, K is the cardinality of \mathbf{x} and \mathbf{e} is the additive noise.

Output: $\hat{\mathbf{x}}_{CoSaMP}$: K -sparse approximation of \mathbf{x}

Initialize the support: $T^0 = \emptyset$.

Initialize the residual: $\mathbf{y}_r^0 = \mathbf{y}$.

while halting criterion is not satisfied **do**

Find new support elements: $T_\Delta = \text{supp}(\mathbf{D}^* \mathbf{y}_r^{\ell-1}, 2K)$.

Update the support: $\tilde{T}^\ell = T^{\ell-1} \cup T_\Delta$.

Compute the representation: $\mathbf{x}_p = \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{y}$.

Prune small entries in the representation: $T^\ell = \text{supp}(\mathbf{x}_p, K)$.

Update the residual: $\mathbf{y}_r^\ell = \mathbf{y} - \mathbf{D}_{T^\ell}(\mathbf{x}_p)_{T^\ell}$.

end while

Form the final solution: $\hat{\mathbf{x}}_{CoSaMP, (T^\ell)^c} = 0$ and $\hat{\mathbf{x}}_{CoSaMP, T^\ell} = (\mathbf{x}_p)_{T^\ell}$.

In the analysis of the CoSaMP that comes next, we follow the same steps as for the SP to derive a near-oracle performance guarantee. Since the proofs are very similar to those of the SP, and those found in [13], we omit most of the derivations and present only the differences.

Theorem 3.4: The CoSaMP solution at the ℓ -th iteration satisfies the recurrence inequality⁶

$$\begin{aligned} \left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 &\leq \frac{4\delta_{4K}}{(1 - \delta_{4K})^2} \left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^{\ell-1} \right\|_2 \\ &\quad + \frac{14 - 6\delta_{4K}}{(1 - \delta_{4K})^2} \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2 \end{aligned} \quad (\text{III.12})$$

For $\delta_{4K} \leq 0.1$ this leads to

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 \leq 0.5 \left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^{\ell-1} \right\|_2 + 16.6 \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \quad (\text{III.13})$$

Proof: The proof of the inequality in (III.12) is given in Appendix D. In a similar way to the proof in the SP case, under the condition $\delta_{4K} \leq 0.1$, it holds that the coefficient multiplying $\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^{\ell-1} \right\|_2$ is smaller or equal to 0.5, while the coefficient multiplying $\left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2$ is smaller or equal to 16.6, which completes our proof. \square

Corollary 3.5: Under the condition $\delta_{4K} \leq 0.1$, the CoSaMP algorithm satisfies

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 \leq 2^{-\ell} \left\| \mathbf{x} \right\|_2 + 2 \cdot 16.6 \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \quad (\text{III.14})$$

In addition, After at most

$$\ell^* = \left\lceil \log_2 \left(\frac{\left\| \mathbf{x} \right\|_2}{\left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2} \right) \right\rceil \quad (\text{III.15})$$

iterations, the solution $\hat{\mathbf{x}}_{CoSaMP}$ leads to an accuracy

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 \leq C_{CoSaMP} \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2, \quad (\text{III.16})$$

where

$$C_{CoSaMP} = \frac{29 - 14\delta_{4K} + \delta_{4K}^2}{(1 - \delta_{4K})^2} \leq 34.1. \quad (\text{III.17})$$

Proof: Starting with (III.13), and applying it recursively, in the same way as was done in the proof of

⁶The observant reader will notice a delicate difference in terminology between this theorem and Theorem 3.1. While here the recurrence formula is expressed with respect to the estimation error, $\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2$, Theorem 3.1 uses a slightly different error measure, $\left\| \mathbf{x}_{T-T^\ell} \right\|_2$.

Corollary 3.5, we obtain

$$\begin{aligned} \left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 &\leq 0.5^k \left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^{\ell-k} \right\|_2 \\ &\quad + 16.6 \cdot \left(\sum_{j=0}^{k-1} 0.5^j \right) \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \end{aligned} \quad (\text{III.18})$$

Setting $k = \ell$ leads easily to (III.14), since $\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^0 \right\|_2 = \left\| \mathbf{x} \right\|_2$.

Plugging the number of iterations ℓ^* as in (III.15) to (III.14) yields⁷

$$\begin{aligned} \left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 &\leq 2^{-\ell^*} \left\| \mathbf{x} \right\|_2 + 2 \cdot \frac{14 - 6\delta_{4K}}{(1 - \delta_{4K})^2} \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2 \\ &\leq \left(1 + 2 \cdot \frac{14 - 6\delta_{4K}}{(1 - \delta_{4K})^2} \right) \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2 \\ &\leq \frac{29 - 14\delta_{4K} + \delta_{4K}^2}{(1 - \delta_{4K})^2} \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \end{aligned}$$

Applying the condition $\delta_{4K} \leq 0.1$ on this equation leads to the result in (III.16). \square

As for the SP, we move now to the random noise case, which leads to a near-oracle performance guarantee for the CoSaMP algorithm.

Theorem 3.6: Assume that \mathbf{e} is a white Gaussian noise vector with variance σ^2 and that the columns of \mathbf{D} are normalized. If the condition $\delta_{4K} \leq 0.1$ holds, then with probability exceeding $1 - (\sqrt{\pi(1+a)} \log N \cdot N^a)^{-1}$ we obtain

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP} \right\|_2^2 \leq C_{CoSaMP}^2 \cdot (2(1+a) \log N) \cdot K \sigma^2. \quad (\text{III.19})$$

Proof: The proof is identical to the one of Theorem 3.6. \square

Fig. 2 shows a graph of C_{CoSaMP} as a function of δ_{4K} . In order to compare the CoSaMP to SP, we also introduce in this figure a graph of C_{SP} versus δ_{4K} (replacing δ_{3K}). Since $\delta_{3K} \leq \delta_{4K}$, the constant C_{SP} is actually better than the values shown in the graph, and yet, it can be seen that even in this case we get $C_{SP} < C_{CoSaMP}$. In addition, the requirement for the SP is expressed with respect to δ_{3K} , while the requirement for the CoSaMP is stronger and uses δ_{4K} .

With comparison to the results presented in [21] for the OMP and the thresholding, the results obtained for the CoSaMP and SP are uniform, expressed only with respect to the properties of the dictionary \mathbf{D} . These algorithms' validity is not dependent on the values of the input vector \mathbf{x} , its energy, or the noise

⁷As before, we replace the constant 16.6 with the equivalent expression that depends on δ_{4K} – see (III.12).

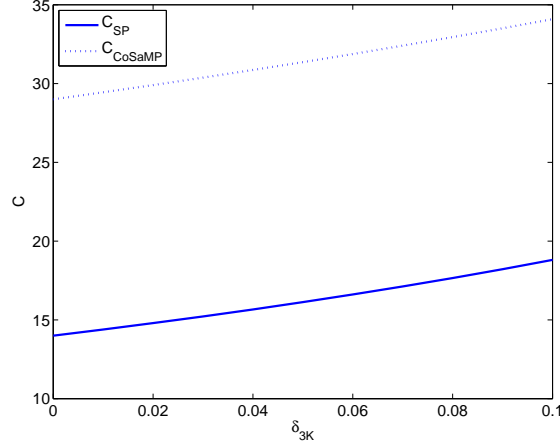


Fig. 2. The constants of the SP and CoSaMP algorithms as a function of δ_{3K}

power. The condition used is the RIP, which implies constraints only on the used dictionary and the sparsity level.

C. Near oracle performance of the IHT algorithm

The IHT algorithm, presented in Algorithm 3, uses a different strategy than the SP and the CoSaMP. It applies only multiplications by \mathbf{D} and \mathbf{D}^* , and a hard thresholding operator. In each iteration it calculates a new representation and keeps its K largest elements. As for the SP and CoSaMP, here as well we employ a fixed number of iterations as a stopping criterion.

Algorithm 3 IHT Algorithm [Equation 7 in [15]]

Input: $K, \mathbf{D}, \mathbf{y}$ where $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e}$, K is the cardinality of \mathbf{x} and \mathbf{e} is the additive noise.

Output: $\hat{\mathbf{x}}_{IHT}$: K -sparse approximation of \mathbf{x}

Initialize the support: $T^0 = \emptyset$.

Initialize the representation: $\mathbf{x}_{IHT}^0 = 0$.

while halting criterion is not satisfied **do**

 Compute the representation: $\mathbf{x}_p = \hat{\mathbf{x}}_{IHT}^{\ell-1} + \mathbf{D}^*(\mathbf{y} - \mathbf{D}\hat{\mathbf{x}}_{IHT}^{\ell-1})$.

 Prune small entries in the representation: $T^\ell = \text{supp}(\mathbf{x}_p, K)$.

 Update the representation: $\hat{\mathbf{x}}_{IHT, (T^\ell)^c}^\ell = 0$ and $\hat{\mathbf{x}}_{IHT, T^\ell}^\ell = (\mathbf{x}_p)_{T^\ell}$.

end while

Form the final solution: $\hat{\mathbf{x}}_{IHT, (T^\ell)^c} = 0$ and $\hat{\mathbf{x}}_{IHT, T^\ell} = (\mathbf{x}_p)_{T^\ell}$.

Similar results, as of the SP and CoSaMP methods, can be sought for the IHT method. Again, the proofs are very similar to the ones shown before for the SP and the CoSaMP and thus only the differences will be presented.

Theorem 3.7: The IHT solution at the ℓ -th iteration satisfies the recurrence inequality

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell \right\|_2 \leq \sqrt{8}\delta_{3K} \left\| \mathbf{x} - \hat{\mathbf{x}}_{IHT}^{\ell-1} \right\|_2 + 4 \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \quad (\text{III.20})$$

For $\delta_{3K} \leq \frac{1}{\sqrt{32}}$ this leads to

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell \right\|_2 \leq 0.5 \left\| \mathbf{x} - \hat{\mathbf{x}}_{IHT}^{\ell-1} \right\|_2 + 4 \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \quad (\text{III.21})$$

Proof: Our proof is based on the proof of Theorem 5 in [15], and only major modifications in the proof will be presented here. Using the definition $\mathbf{r}^\ell \triangleq \mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell$, and an inequality taken from Equation (22) in [15], it holds that

$$\begin{aligned} \left\| \mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell \right\|_2 &\leq 2 \left\| \mathbf{x}_{T \cup T^\ell} - (\mathbf{x}_p)_{T \cup T^\ell} \right\|_2 \\ &= 2 \left\| \mathbf{x}_{T \cup T^\ell} - \hat{\mathbf{x}}_{T \cup T^\ell}^{\ell-1} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D} \mathbf{r}^{\ell-1} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{e} \right\|_2 \\ &= 2 \left\| \mathbf{r}_{T \cup T^\ell}^{\ell-1} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D} \mathbf{r}^{\ell-1} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{e} \right\|_2, \end{aligned} \quad (\text{III.22})$$

where the equality emerges from the definition $\mathbf{x}_p = \hat{\mathbf{x}}_{IHT}^{\ell-1} + \mathbf{D}^*(\mathbf{y} - \mathbf{D}\hat{\mathbf{x}}_{IHT}^{\ell-1}) = \hat{\mathbf{x}}_{IHT}^{\ell-1} + \mathbf{D}^*(\mathbf{D}\mathbf{x} + \mathbf{e} - \mathbf{D}\hat{\mathbf{x}}_{IHT}^{\ell-1})$.

The support of $\mathbf{r}^{\ell-1}$ is over $T \cup T^{\ell-1}$ and thus it is also over $T \cup T^\ell \cup T^{\ell-1}$. Based on this, we can divide $\mathbf{D}\mathbf{r}^{\ell-1}$ into a part supported on $T^{\ell-1} - T^\ell \cup T$ and a second part supported on $T^\ell \cup T$. Using this and the triangle inequality with (III.22), we obtain

$$\begin{aligned} &\left\| \mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell \right\|_2 \\ &\leq 2 \left\| \mathbf{r}_{T \cup T^\ell}^{\ell-1} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D} \mathbf{r}^{\ell-1} \right\|_2 + 2 \left\| \mathbf{D}_{T \cup T^\ell}^* \mathbf{e} \right\|_2 \\ &= 2 \left\| (\mathbf{I} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D}_{T \cup T^\ell}) \mathbf{r}_{T \cup T^\ell}^{\ell-1} \right. \\ &\quad \left. - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D}_{T^{\ell-1} - T \cup T^\ell} \mathbf{r}_{T^{\ell-1} - T \cup T^\ell}^{\ell-1} \right\|_2 + 2 \left\| \mathbf{D}_{T \cup T^\ell}^* \mathbf{e} \right\|_2 \\ &\leq 2 \left\| (\mathbf{I} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D}_{T \cup T^\ell}) \mathbf{r}_{T \cup T^\ell}^{\ell-1} \right\|_2 \\ &\quad + 2 \left\| \mathbf{D}_{T \cup T^\ell}^* \mathbf{D}_{T^{\ell-1} - T \cup T^\ell} \mathbf{r}_{T^{\ell-1} - T \cup T^\ell}^{\ell-1} \right\|_2 + 2 \left\| \mathbf{D}_{T_e, 2}^* \mathbf{e} \right\|_2 \\ &\leq 2\delta_{2K} \left\| \mathbf{r}_{T \cup T^\ell}^{\ell-1} \right\|_2 + 2\delta_{3K} \left\| \mathbf{r}_{T^{\ell-1} - T \cup T^\ell}^{\ell-1} \right\|_2 + 4 \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2. \end{aligned} \quad (\text{III.23})$$

The last inequality holds because the eigenvalues of $(\mathbf{I} - \mathbf{D}_{T \cup T^\ell}^* \mathbf{D}_{T \cup T^\ell})$ are in the range $[-\delta_{2K}, \delta_{2K}]$, the size of the set $T \cup T^\ell$ is smaller than $2K$, the sets $T \cup T^\ell$ and $T^{\ell-1} - T \cup T^\ell$ are disjoint, and of total size of these together is equal or smaller than $3K$. Note that we have used the definition of $T_{e,2}$ as

given in Section II.

We proceed by observing that $\|\mathbf{r}_{T^{\ell-1}-T \cup T^\ell}^{\ell-1}\|_2 + \|\mathbf{r}_{T \cup T^\ell}^{\ell-1}\|_2 \leq \sqrt{2} \|\mathbf{r}^{\ell-1}\|_2$, since these vectors are orthogonal. Using the fact that $\delta_{2K} \leq \delta_{3K}$ we get (III.20) from (III.23). Finally, under the condition $\delta_{3K} \leq 1/\sqrt{32}$, it holds that the coefficient multiplying $\|\mathbf{x} - \hat{\mathbf{x}}_{IHT}^{\ell-1}\|_2$ is smaller or equal to 0.5, which completes our proof. \square

Corollary 3.8: Under the condition $\delta_{3K} \leq 1/\sqrt{32}$, the IHT algorithm satisfies

$$\|\mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell\|_2 \leq 2^{-\ell} \|\mathbf{x}\|_2 + 8 \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \quad (\text{III.24})$$

In addition, After at most

$$\ell^* = \left\lceil \log_2 \left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2} \right) \right\rceil \quad (\text{III.25})$$

iterations, the solution $\hat{\mathbf{x}}_{IHT}$ leads to an accuracy

$$\|\mathbf{x} - \hat{\mathbf{x}}_{IHT}^\ell\|_2 \leq C_{IHT} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2, \quad (\text{III.26})$$

where

$$C_{IHT} = 9. \quad (\text{III.27})$$

Proof: The proof is obtained following the same steps as in Corollaries 3.2 and 3.5. \square

Finally, considering a random noise instead of an adversarial one, we get a near-oracle performance guarantee for the IHT algorithm, as was achieved for the SP and CoSaMP.

Theorem 3.9: Assume that \mathbf{e} is a white Gaussian noise with variance σ^2 and that the columns of \mathbf{D} are normalized. If the condition $\delta_{3K} \leq 1/\sqrt{32}$ holds, then with probability exceeding $1 - (\sqrt{\pi(1+a)} \log N \cdot N^a)^{-1}$ we obtain

$$\|\mathbf{x} - \hat{\mathbf{x}}_{IHT}\|_2^2 \leq C_{IHT}^2 \cdot (2(1+a) \log N) \cdot K \sigma^2. \quad (\text{III.28})$$

Proof: The proof is identical to the one of Theorem 3.6. \square

A comparison between the constants achieved by the IHT, SP and DS is presented in Fig. 3. The CoSaMP constant was omitted since it is bigger than the one of the SP and it is dependent on δ_{4K} instead of δ_{3K} . The figure shows that the constant values of IHT and DS are better than that of the SP (and as such better than the one of the CoSaMP), and that the one of the DS is the smallest. It is interesting to note that the constant of the IHT is independent of δ_{3K} .

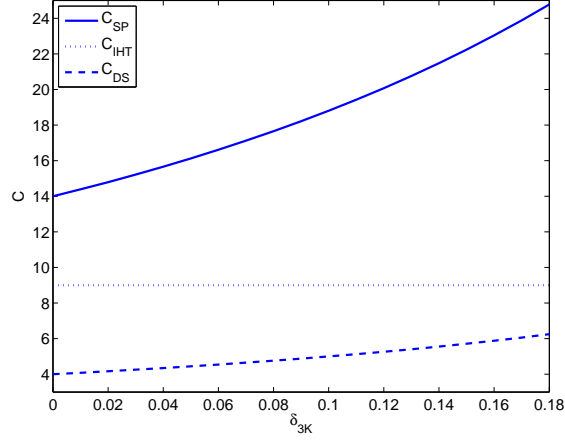


Fig. 3. The constants of the SP, IHT and DS algorithms as a function of δ_{3K}

Alg.	RIP Condition	Probability of Correctness	Constant	The Obtained Bound
DS	$\delta_{2K} + \delta_{3K} \leq 1$	$1 - (\sqrt{\pi(1+a) \log N} \cdot N^a)^{-1}$	$\frac{4}{1-2\delta_{3K}}$	$C_{DS}^2 \cdot (2(1+a) \log N) \cdot K\sigma^2$
BP	$\delta_{2K} + 3\delta_{3K} \leq 1$	$1 - (N^a)^{-1}$	$> \frac{32}{\kappa^4}$	$C_{BP}^2 \cdot (2(1+a) \log N) \cdot K\sigma^2$
SP	$\delta_{3K} \leq 0.139$	$1 - (\sqrt{\pi(1+a) \log N} \cdot N^a)^{-1}$	≤ 21.41	$C_{SP}^2 \cdot (2(1+a) \log N) \cdot K\sigma^2$
CoSaMP	$\delta_{4K} \leq 0.1$	$1 - (\sqrt{\pi(1+a) \log N} \cdot N^a)^{-1}$	≤ 34.2	$C_{CoSaMP}^2 \cdot (2(1+a) \log N) \cdot K\sigma^2$
IHT	$\delta_{3K} \leq \frac{1}{\sqrt{32}}$	$1 - (\sqrt{\pi(1+a) \log N} \cdot N^a)^{-1}$	9	$C_{IHT}^2 \cdot (2(1+a) \log N) \cdot K\sigma^2$

TABLE I
NEAR ORACLE PERFORMANCE GUARANTEES FOR THE DS, BP, SP, CoSaMP AND IHT TECHNIQUES.

In table I we summarize the performance guarantees of several different algorithms – the DS [8], the BP [29], and the three algorithms analyzed in this paper.

We can observe the following:

- 1) In terms of the RIP: DS and BP are the best, then IHT, then SP and last is CoSaMP.
- 2) In terms of the constants in the bounds: the smallest constant is achieved by DS. Then come IHT, SP, CoSaMP and BP in this order.
- 3) In terms of the probability: all have the same probability except the BP which gives a weaker guarantee.
- 4) Though the CoSaMP has a weaker guarantee compared to the SP, it has an efficient implementation that saves the matrix inversion in the algorithm.⁸

⁸The proofs of the guarantees in this paper are not valid for this case, though it is not hard to extend them for it.

For completeness of the discussion here, we also refer to algorithms' complexity: the IHT is the cheapest, CoSaMP and SP come next with a similar complexity (with a slight advantage to CoSaMP). DS and BP seem to be the most complex.

Interestingly, in the guarantees of the OMP and the thresholding in [21] better constants are obtained. However, these results, as mentioned before, holds under mutual-coherence based conditions, which are more restricting. In addition, their validity relies on the magnitude of the entries of \mathbf{x} and the noise power, which is not correct for the results presented in this section for the greedy-like methods. Furthermore, though we get bigger constants with these methods, the conditions are not tight, as will be seen in the next section.

IV. EXPERIMENTS

In our experiments we use a random dictionary with entries drawn from the canonic normal distribution. The columns of the dictionary are normalized and the dimensions are $m = 512$ and $N = 1024$. The vector \mathbf{x} is generated by selecting a support uniformly at random. Then the elements in the support are generated using the following model⁹:

$$\mathbf{x}_i = 10\epsilon_i(1 + |n_i|) \quad (\text{IV.1})$$

where ϵ_i is ± 1 with probability 0.5, and n_i is a canonic normal random variable. The support and the non-zero values are statistically independent. We repeat each experiment 1500 times.

In the first experiment we calculate the error of the SP, CoSaMP and IHT methods for different sparsity levels. The noise variance is set to $\sigma = 1$. Fig. 4 presents the squared-error $\|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ of all the instances of the experiment for the three algorithms. Our goal is to show that with high-probability the error obtained is bounded by the guarantees we have developed. For each algorithm we add the theoretical guarantee and the oracle performance. As can be seen, the theoretical guarantees are too loose and the actual performance of the algorithms is much better. However, we see that both the theoretical and the empirical performance curves show a proportionality to the oracle error. Note that the actual performance of the algorithms' may be better than the oracle's – this happens because the oracle is the Maximum-Likelihood Estimator (MLE) in this case [31], and by adding a bias one can perform even better in some cases.

⁹This model is taken from the experiments section in [8].

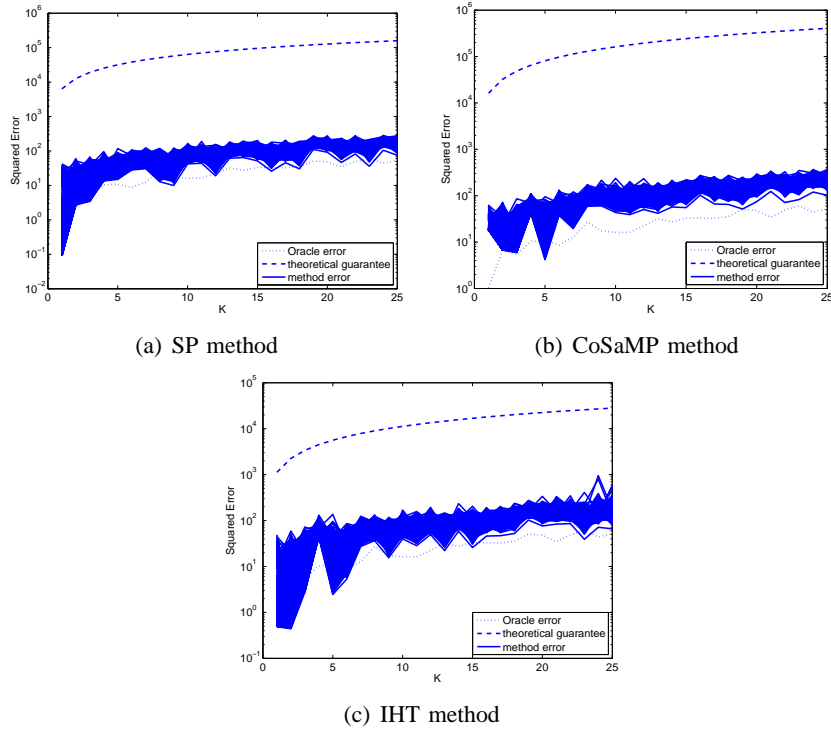


Fig. 4. The squared-error as achieved by the SP, the CoSaMP and the IHT algorithms as a function of the cardinality. The graphs also show the theoretical guarantees and the oracle performance.

Fig. 5(a) presents the mean-squared-error (by averaging all the experiments) for the range where the RIP-condition seems to hold, and Fig. 5(b) presents this error for a wider range, where it is likely to be violated. It can be seen that in the average case, though the algorithms get different constants in their bounds, they achieve almost the same performance. We also see a near-linear curve describing the error as a function of K . Finally, we observe that the SP and the CoSaMP, which were shown to have worse constants in theory, have better performance and are more stable in the case where the RIP-condition does not hold anymore.

In a second experiment we calculate the error of the SP, the CoSaMP and the IHT methods for different noise variances. The sparsity is set to $K = 10$. Fig. 6 presents the error of all the instances of the experiment for the three algorithms. Here as well we add the theoretical guarantee and the oracle performance. As we saw before, the guarantee is not tight but the error is proportional to the oracle estimator's error.

Fig. 7 presents the mean-squared-error as a function of the noise variance, by averaging over all the experiments. It can be seen that the error behaves linearly with respect to the variance, as expected from the theoretical analysis. Again we see that the constants are not tight and that the algorithms behave

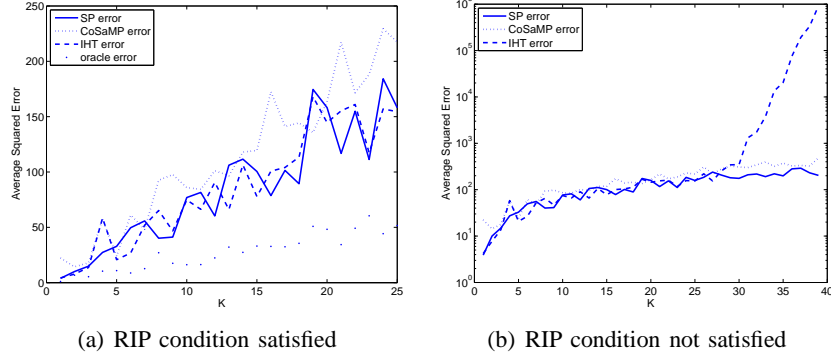


Fig. 5. The mean-squared-error of the SP, the CoSaMP and the IHT algorithms as a function of the cardinality.

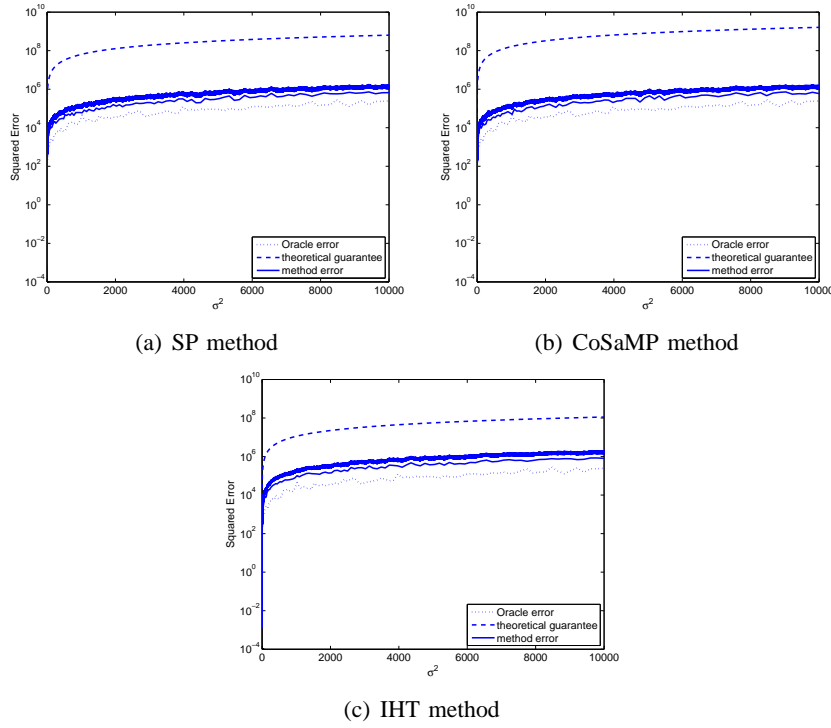


Fig. 6. The squared-error as achieved by the SP, the CoSaMP and the IHT algorithms as a function of the noise variance. The graphs also show the theoretical guarantees and the oracle performance.

in a similar way. Finally, we note that the algorithms succeed in meeting the bounds even in very low signal-to-noise ratios, where simple greedy algorithms are expected to fail.

V. EXTENSION TO THE NON-EXACT SPARSE CASE

In the case where \mathbf{x} is not exactly K -sparse, our analysis has to change. Following the work reported in [13], we have the following error bounds for all algorithms (with the different RIP condition and

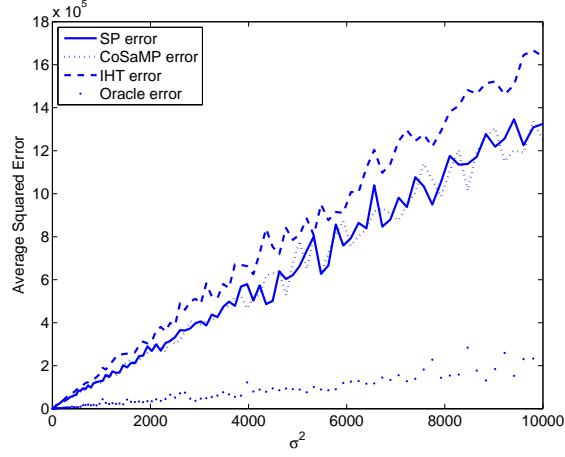


Fig. 7. The mean-squared-error of the SP, the CoSaMP and the IHT algorithms as a function of the noise variance.

constant).

Theorem 5.1: For the SP, CoSaMP and IHT algorithms, under their appropriate RIP conditions, it holds that after at most

$$\ell^* = \left\lceil \log_2 \left(\frac{\|\mathbf{x}\|_2}{\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2} \right) \right\rceil \quad (\text{V.1})$$

iterations, the estimation $\hat{\mathbf{x}}$ gives an accuracy of the form

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}\|_2 \leq C \bigg(& \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\ & + (1 + \delta_K) \|\mathbf{x} - \mathbf{x}_K\|_2 + \frac{1 + \delta_K}{\sqrt{K}} \|\mathbf{x} - \mathbf{x}_K\|_1 \bigg). \end{aligned} \quad (\text{V.2})$$

where \mathbf{x}_K is a K -sparse vector that nulls all entries in \mathbf{x} apart from the K dominant ones. C is the appropriate constant value, dependent on the algorithm.

If we assume that \mathbf{e} is a white Gaussian noise with variance σ^2 and that the columns of \mathbf{D} are normalized, then with probability exceeding $1 - (\sqrt{\pi(1+a)\log N} \cdot N^a)^{-1}$ we get that

$$\begin{aligned} \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 \leq 2 \cdot C^2 \bigg(& \sqrt{(1+a)\log N} \cdot K \cdot \sigma \\ & + \|\mathbf{x} - \mathbf{x}_K\|_2 + \frac{1}{\sqrt{K}} \|\mathbf{x} - \mathbf{x}_K\|_1 \bigg)^2. \end{aligned} \quad (\text{V.3})$$

Proof: Proposition 3.5 from [13] provides us with the following claim

$$\|\mathbf{D}\mathbf{x}\|_2 \leq \sqrt{1 + \delta_K} \left(\|\mathbf{x}\|_2 + \frac{1}{\sqrt{K}} \|\mathbf{x}\|_1 \right). \quad (\text{V.4})$$

When \mathbf{x} is not exactly K -sparse we get that the effective error in our results becomes $\tilde{\mathbf{e}} = \mathbf{e} + \mathbf{D}(\mathbf{x} - \mathbf{x}_K)$. Thus, using the error bounds of the algorithms with the inequality in (V.4) we get

$$\begin{aligned}
 \|\mathbf{x} - \hat{\mathbf{x}}\|_2 &\leq C \|\mathbf{D}_{T_e}^* \tilde{\mathbf{e}}\|_2 \\
 &\leq C \|\mathbf{D}_{T_e}^* (\mathbf{e} + \mathbf{D}(\mathbf{x} - \mathbf{x}_K))\|_2 \\
 &\leq C \|\mathbf{D}_{T_e}^* \mathbf{e}\| + C \|\mathbf{D}_{T_e}^* \mathbf{D}(\mathbf{x} - \mathbf{x}_K)\|_2 \\
 &\leq C \|\mathbf{D}_{T_e}^* \mathbf{e}\| + C \sqrt{1 + \delta_K} \|\mathbf{D}(\mathbf{x} - \mathbf{x}_K)\|_2 \\
 &\leq C \left(\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 + (1 + \delta_K) \|\mathbf{x} - \mathbf{x}_K\|_2 \right. \\
 &\quad \left. + \frac{1 + \delta_K}{\sqrt{K}} \|\mathbf{x} - \mathbf{x}_K\|_1 \right),
 \end{aligned} \tag{V.5}$$

which proves (V.2). Using the same steps taken in Theorems 3.3, 3.6, and 3.9, lead us to

$$\begin{aligned}
 \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 &\leq C^2 \left(\sqrt{(2(1 + a) \log N) \cdot K} \cdot \sigma \right. \\
 &\quad \left. + (1 + \delta_K) \|\mathbf{x} - \mathbf{x}_K\|_2 + \frac{1 + \delta_K}{\sqrt{K}} \|\mathbf{x} - \mathbf{x}_K\|_1 \right)^2.
 \end{aligned} \tag{V.6}$$

Since the RIP condition for all the algorithms satisfies $\delta_K \leq \sqrt{2} - 1$, plugging this into (V.6) gives (V.3), and this concludes the proof. \square

Just as before, we should wonder how close is this bound to the one obtained by an oracle that knows the support T of the K dominant entries in \mathbf{x} . Following [15], we derive an expression for such an oracle. Using the fact that the oracle is given by $\mathbf{D}_T^\dagger \mathbf{y} = \mathbf{D}_T^\dagger (\mathbf{D} \mathbf{x}_T + \mathbf{D}(\mathbf{x} - \mathbf{x}_T) + \mathbf{e})$, its MSE is bounded by

$$\begin{aligned}
 E \|\mathbf{x} - \hat{\mathbf{x}}_{oracle}\|_2^2 &= E \left\| \mathbf{x} - \mathbf{D}_T^\dagger \mathbf{y} \right\|_2^2 \\
 &= E \left\| \mathbf{x} - \mathbf{x}_T - \mathbf{D}_T^\dagger \mathbf{e} - \mathbf{D}_T^\dagger \mathbf{D}(\mathbf{x} - \mathbf{x}_T) \right\|_2^2 \\
 &\leq \left(\|\mathbf{x} - \mathbf{x}_T\|_2 + \left\| \mathbf{D}_T^\dagger \mathbf{D}(\mathbf{x} - \mathbf{x}_T) \right\|_2 + E \left\| \mathbf{D}_T^\dagger \mathbf{e} \right\|_2 \right)^2,
 \end{aligned} \tag{V.7}$$

where we have used the triangle inequality. Using the relation given in (I.10) for the last term, and properties of the RIP for the second, we obtain

$$\begin{aligned}
 E \|\mathbf{x} - \hat{\mathbf{x}}_{oracle}\|_2^2 &\leq \\
 &\left(\|\mathbf{x} - \mathbf{x}_T\|_2 + \frac{1}{\sqrt{1 - \delta_K}} \|\mathbf{D}(\mathbf{x} - \mathbf{x}_T)\|_2 + \frac{\sqrt{K}}{\sqrt{1 - \delta_K}} \sigma \right)^2.
 \end{aligned} \tag{V.8}$$

Finally, the middle-term can be further handled using (V.4), and we arrive to

$$E \|\mathbf{x} - \hat{\mathbf{x}}_{oracle}\|_2^2 \leq \frac{1}{1 - \delta_k} \left((1 + \sqrt{1 + \delta_K}) \|\mathbf{x} - \mathbf{x}_T\|_2 + \frac{\sqrt{1 + \delta_K}}{\sqrt{K}} \|\mathbf{x} - \mathbf{x}_K\|_1 + \sqrt{K} \sigma \right)^2. \quad (\text{V.9})$$

Thus we see again that the error bound in the non-exact sparse case, is up to a constant and the $\log N$ factor the same as the one of the oracle estimator.

VI. CONCLUSION

In this paper we have presented near-oracle performance guarantees for three greedy-like algorithms – the Subspace Pursuit, the CoSaMP, and the Iterative Hard-Thresholding. The approach taken in our analysis is an RIP-based (as opposed to mutual-coherence ones). Despite their resemblance to greedy algorithms, such as the OMP and the thresholding, our study leads to uniform guarantees for the three algorithms explored, i.e., the near-oracle error bounds are dependent only on the dictionary properties (RIP constant) and the sparsity level of the sought solution. We have also presented a simple extension of our results to the case where the representations are only approximately sparse.

ACKNOWLEDGMENT

The authors would like to thank Zvika Ben-Haim for fruitful discussions and relevant references to existing literature, which helped in shaping this work.

APPENDIX A

PROOF OF THEOREM 3.1 – INEQUALITY (III.1)

In the proof of (III.1) we use two main inequalities:

$$\begin{aligned} \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 &\leq \frac{2\delta_{3K}}{(1 - \delta_{3K})^2} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 \\ &\quad + \frac{2}{(1 - \delta_{3K})^2} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2, \end{aligned} \quad (\text{A.1})$$

and

$$\begin{aligned} \|\mathbf{x}_{T-T^\ell}\|_2 &\leq \frac{1 + \delta_{3K}}{1 - \delta_{3K}} \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 \\ &\quad + \frac{4}{1 - \delta_{3K}} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{A.2})$$

Their proofs are in Appendices B and C respectively. The inequality (III.1) is obtained by substituting (A.1) into (A.2) as shown below:

$$\begin{aligned}
\|\mathbf{x}_{T-T^\ell}\|_2 &\leq \frac{1+\delta_{3K}}{1-\delta_{3K}} \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 + \frac{4}{1-\delta_{3K}} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\
&\leq \frac{1+\delta_{3K}}{1-\delta_{3K}} \left[\frac{2\delta_{3K}}{(1-\delta_{3K})^2} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 \right. \\
&\quad \left. + \frac{2}{(1-\delta_{3K})^2} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \right] + \frac{4}{1-\delta_{3K}} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\
&\leq \frac{2\delta_{3K}(1+\delta_{3K})}{(1-\delta_{3K})^3} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 \\
&\quad + \frac{2(1+\delta_{3K})}{(1-\delta_{3K})^3} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 + \frac{4}{1-\delta_{3K}} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2 \\
&\leq \frac{2\delta_{3K}(1+\delta_{3K})}{(1-\delta_{3K})^3} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 \\
&\quad + \frac{6-6\delta_{3K}+4\delta_{3K}^2}{(1-\delta_{3K})^3} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2,
\end{aligned} \tag{A.3}$$

and this concludes this proof. \square

APPENDIX B

PROOF OF INEQUALITY (A.1)

Lemma B.1: The following inequality holds true for the SP algorithm:

$$\begin{aligned}
\|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 &\leq \frac{2\delta_{3K}}{(1-\delta_{3K})^2} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 \\
&\quad + \frac{2}{(1-\delta_{3K})^2} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2,
\end{aligned}$$

Proof: We start by the residual-update step in the SP algorithm, and exploit the relation $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e} = \mathbf{D}_{T-T^{\ell-1}}\mathbf{x}_{T-T^{\ell-1}} + \mathbf{D}_{T\cap T^{\ell-1}}\mathbf{x}_{T\cap T^{\ell-1}} + \mathbf{e}$. This leads to

$$\begin{aligned}
\mathbf{y}_r^{\ell-1} &= \text{resid}(\mathbf{y}, \mathbf{D}_{T^{\ell-1}}) \\
&= \text{resid}(\mathbf{D}_{T-T^{\ell-1}}\mathbf{x}_{T-T^{\ell-1}}, \mathbf{D}_{T^{\ell-1}}) \\
&\quad + \text{resid}(\mathbf{D}_{T\cap T^{\ell-1}}\mathbf{x}_{T\cap T^{\ell-1}}, \mathbf{D}_{T^{\ell-1}}) + \text{resid}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}}).
\end{aligned} \tag{B.1}$$

Here we have used the linearity of the operator $\text{resid}(\cdot, \mathbf{D}_{T^{\ell-1}})$ with respect to its first entry. The second term in the right-hand-side (rhs) is 0 since $\mathbf{D}_{T\cap T^{\ell-1}}\mathbf{x}_{T\cap T^{\ell-1}} \in \text{span}(\mathbf{D}_{T^{\ell-1}})$. For the first term in the

rhs we have

$$\begin{aligned}
& \text{resid}(\mathbf{D}_{T-T^{\ell-1}} \mathbf{x}_{T-T^{\ell-1}}, \mathbf{D}_{T^{\ell-1}}) \\
&= \mathbf{D}_{T-T^{\ell-1}} \mathbf{x}_{T-T^{\ell-1}} - \text{proj}(\mathbf{D}_{T-T^{\ell-1}} \mathbf{x}_{T-T^{\ell-1}}, \mathbf{D}_{T^{\ell-1}}) \\
&= \mathbf{D}_{T-T^{\ell-1}} \mathbf{x}_{T-T^{\ell-1}} + \mathbf{D}_{T^{\ell-1}} \mathbf{x}_{p, T^{\ell-1}} \\
&= [\mathbf{D}_{T-T^{\ell-1}}, \mathbf{D}_{T^{\ell-1}}] \begin{bmatrix} \mathbf{x}_{T-T^{\ell-1}} \\ \mathbf{x}_{p, T^{\ell-1}} \end{bmatrix} \triangleq \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1},
\end{aligned} \tag{B.2}$$

where we have defined

$$\mathbf{x}_{p, T^{\ell-1}} = -(\mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T^{\ell-1}})^{-1} \mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T-T^{\ell-1}} \mathbf{x}_{T-T^{\ell-1}}. \tag{B.3}$$

Combining (B.1) and (B.2) leads to

$$\mathbf{y}_r^{\ell-1} = \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} + \text{resid}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}}). \tag{B.4}$$

By the definition of T_Δ in Algorithm 1 we obtain

$$\begin{aligned}
& \|\mathbf{D}_{T_\Delta}^* \mathbf{y}_r^{\ell-1}\|_2 \geq \|\mathbf{D}_T^* \mathbf{y}_r^{\ell-1}\|_2 \geq \|\mathbf{D}_{T-T^{\ell-1}}^* \mathbf{y}_r^{\ell-1}\|_2 \\
& \geq \|\mathbf{D}_{T-T^{\ell-1}}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1}\|_2 \\
& \quad - \|\mathbf{D}_{T-T^{\ell-1}}^* \text{resid}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}})\|_2 \\
& \geq \|\mathbf{D}_{T-T^{\ell-1}}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1}\|_2 - \|\mathbf{D}_{T-T^{\ell-1}}^* \mathbf{e}\|_2 \\
& \quad - \|\mathbf{D}_{T-T^{\ell-1}}^* \text{proj}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}})\|_2.
\end{aligned} \tag{B.5}$$

We will bound $\|\mathbf{D}_{T-T^{\ell-1}}^* \text{proj}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}})\|_2$ from above using RIP properties from Section II,

$$\begin{aligned}
& \|\mathbf{D}_{T-T^{\ell-1}}^* \text{proj}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}})\|_2 \\
&= \|\mathbf{D}_{T-T^{\ell-1}}^* \mathbf{D}_{T^{\ell-1}} (\mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T^{\ell-1}})^{-1} \mathbf{D}_{T^{\ell-1}}^* \mathbf{e}\|_2 \\
&\leq \frac{\delta_{3K}}{1 - \delta_{3K}} \|\mathbf{D}_{T^{\ell-1}}^* \mathbf{e}\|_2.
\end{aligned} \tag{B.6}$$

Combining (B.5) and (B.6) leads to

$$\begin{aligned}
\left\| \mathbf{D}_{T_\Delta}^* \mathbf{y}_r^{\ell-1} \right\|_2 &\geq \left\| \mathbf{D}_{T-T^{\ell-1}}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 \\
&\quad - \left\| \mathbf{D}_T^* \mathbf{e} \right\|_2 - \frac{\delta_{3K}}{1 - \delta_{3K}} \left\| \mathbf{D}_{T^{\ell-1}}^* \mathbf{e} \right\|_2 \\
&\geq \left\| \mathbf{D}_{T-T^{\ell-1}}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 - \frac{1}{1 - \delta_{3K}} \left\| \mathbf{D}_{T_*}^* \mathbf{e} \right\|_2.
\end{aligned} \tag{B.7}$$

By the definition of T_Δ and $\mathbf{y}_r^{\ell-1}$ it holds that $T_\Delta \cap T^{\ell-1} = \emptyset$ since $\mathbf{D}_{T^{\ell-1}}^* \mathbf{y}_r^{\ell-1} = 0$. Using (B.4), the left-hand-side (lhs) of (B.7) is upper bounded by

$$\begin{aligned}
\left\| \mathbf{D}_{T_\Delta}^* \mathbf{y}_r^{\ell-1} \right\|_2 &\leq \left\| \mathbf{D}_{T_\Delta}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 + \left\| \mathbf{D}_{T_\Delta}^* \text{resid}(\mathbf{e}, \mathbf{D}_{T^{\ell-1}}) \right\|_2 \\
&\leq \left\| \mathbf{D}_{T_\Delta}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 + \left\| \mathbf{D}_{T_\Delta}^* \mathbf{e} \right\|_2 \\
&\quad + \left\| \mathbf{D}_{T_\Delta}^* \mathbf{D}_{T^{\ell-1}} (\mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T^{\ell-1}})^{-1} \mathbf{D}_{T^{\ell-1}}^* \mathbf{e} \right\|_2 \\
&\leq \left\| \mathbf{D}_{T_\Delta}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 + \left\| \mathbf{D}_{T_\Delta}^* \mathbf{e} \right\|_2 \\
&\quad + \frac{\delta_{3K}}{1 - \delta_{3K}} \left\| \mathbf{D}_{T^{\ell-1}}^* \mathbf{e} \right\|_2 \\
&\leq \left\| \mathbf{D}_{T_\Delta}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 + \frac{1}{1 - \delta_{3K}} \left\| \mathbf{D}_{T_*}^* \mathbf{e} \right\|_2.
\end{aligned} \tag{B.8}$$

Combining (B.7) and (B.8) gives

$$\begin{aligned}
&\left\| \mathbf{D}_{T_\Delta}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 + \frac{2}{1 - \delta_{3K}} \left\| \mathbf{D}_{T_*}^* \mathbf{e} \right\|_2 \\
&\geq \left\| \mathbf{D}_{T-T^{\ell-1}}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2.
\end{aligned} \tag{B.9}$$

Removing the common rows in $\mathbf{D}_{T_\Delta}^*$ and $\mathbf{D}_{T-T^{\ell-1}}^*$ we get

$$\begin{aligned}
&\left\| \mathbf{D}_{T_\Delta-T}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 + \frac{2}{1 - \delta_{3K}} \left\| \mathbf{D}_{T_*}^* \mathbf{e} \right\|_2 \\
&\geq \left\| \mathbf{D}_{T-T^{\ell-1}-T_\Delta}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 \\
&= \left\| \mathbf{D}_{T-\tilde{T}^\ell}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2.
\end{aligned} \tag{B.10}$$

The last equality is true because $T - T^{\ell-1} - T_\Delta = T - T^{\ell-1} - (\tilde{T}^\ell - T^{\ell-1}) = T - \tilde{T}^\ell$.

Now we turn to bound the lhs and rhs terms of (B.10) from below and above, respectively. For the lhs term we exploit the fact that the supports $T_\Delta - T$ and $T \cup T^{\ell-1}$ are disjoint, leading to

$$\begin{aligned} \left\| \mathbf{D}_{T_\Delta - T}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 &\leq \delta_{|T_\Delta \cup T^{\ell-1} \cup T|} \left\| \mathbf{x}_r^{\ell-1} \right\|_2 \\ &\leq \delta_{3K} \left\| \mathbf{x}_r^{\ell-1} \right\|_2 \end{aligned} \quad (\text{B.11})$$

For the rhs term in (B.10), we obtain

$$\begin{aligned} &\left\| \mathbf{D}_{T - \tilde{T}^\ell}^* \mathbf{D}_{T \cup T^{\ell-1}} \mathbf{x}_r^{\ell-1} \right\|_2 \\ &\geq \left\| \mathbf{D}_{T - \tilde{T}^\ell}^* \mathbf{D}_{T - \tilde{T}^\ell} (\mathbf{x}_r^{\ell-1})_{T - \tilde{T}^\ell} \right\|_2 \\ &\quad - \left\| \mathbf{D}_{T - \tilde{T}^\ell}^* \mathbf{D}_{(T \cup T^{\ell-1}) - (T - \tilde{T}^\ell)} (\mathbf{x}_r^{\ell-1})_{(T \cup T^{\ell-1}) - (T - \tilde{T}^\ell)} \right\|_2 \\ &\geq (1 - \delta_K) \left\| (\mathbf{x}_r^{\ell-1})_{T - \tilde{T}^\ell} \right\|_2 - \delta_{3K} \left\| \mathbf{x}_r^{\ell-1} \right\|_2 \\ &\geq (1 - \delta_{3K}) \left\| (\mathbf{x}_r^{\ell-1})_{T - \tilde{T}^\ell} \right\|_2 - \delta_{3K} \left\| \mathbf{x}_r^{\ell-1} \right\|_2 \end{aligned} \quad (\text{B.12})$$

Substitution of the two bounds derived above into (B.10) gives

$$\begin{aligned} &2\delta_{3K} \left\| \mathbf{x}_r^{\ell-1} \right\|_2 + \frac{2}{1 - \delta_{3K}} \left\| \mathbf{D}_{T_e}^* \mathbf{e} \right\|_2 \\ &\geq (1 - \delta_{3K}) \left\| (\mathbf{x}_r^{\ell-1})_{T - \tilde{T}^\ell} \right\|_2. \end{aligned} \quad (\text{B.13})$$

The above inequality uses $\mathbf{x}_r^{\ell-1}$, which was defined in (B.2), and this definition relies on yet another one definition for the vector $\mathbf{x}_{p, T^{\ell-1}}$ in (B.3). We proceed by bounding $\left\| \mathbf{x}_{p, T^{\ell-1}} \right\|_2$ from above,

$$\begin{aligned} &\left\| \mathbf{x}_{p, T^{\ell-1}} \right\|_2 \\ &= \left\| -(\mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T^{\ell-1}})^{-1} \mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T - T^{\ell-1}} \mathbf{x}_{T - T^{\ell-1}} \right\|_2 \\ &\leq \frac{1}{1 - \delta_K} \left\| -\mathbf{D}_{T^{\ell-1}}^* \mathbf{D}_{T - T^{\ell-1}} \mathbf{x}_{T - T^{\ell-1}} \right\|_2 \\ &\leq \frac{\delta_{2K}}{1 - \delta_K} \left\| \mathbf{x}_{T - T^{\ell-1}} \right\|_2 \leq \frac{\delta_{3K}}{1 - \delta_{3K}} \left\| \mathbf{x}_{T - T^{\ell-1}} \right\|_2, \end{aligned} \quad (\text{B.14})$$

and get

$$\begin{aligned} \left\| \mathbf{x}_r^{\ell-1} \right\|_2 &\leq \left\| \mathbf{x}_{T - T^{\ell-1}} \right\|_2 + \left\| \mathbf{x}_{p, T^{\ell-1}} \right\|_2 \\ &\leq \left(1 + \frac{\delta_{3K}}{1 - \delta_{3K}} \right) \left\| \mathbf{x}_{T - T^{\ell-1}} \right\|_2 \\ &\leq \frac{1}{1 - \delta_{3K}} \left\| \mathbf{x}_{T - T^{\ell-1}} \right\|_2. \end{aligned} \quad (\text{B.15})$$

In addition, since $(\mathbf{x}_r^{\ell-1})_{T-T^{\ell-1}} = \mathbf{x}_{T-T^{\ell-1}}$ then $(\mathbf{x}_r^{\ell-1})_{T-\tilde{T}^\ell} = \mathbf{x}_{T-\tilde{T}^\ell}$. Using this fact and (B.15) with (B.13) leads to

$$\begin{aligned} & \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 \\ & \leq \frac{2\delta_{3K}}{(1-\delta_{3K})^2} \|\mathbf{x}_{T-T^{\ell-1}}\|_2 + \frac{2}{(1-\delta_{3K})^2} \|\mathbf{D}_{T_*}^* \mathbf{e}\|_2, \end{aligned} \quad (\text{B.16})$$

which proves the inequality in (A.1). \square

APPENDIX C

PROOF OF INEQUALITY (A.2)

Lemma C.1: The following inequality holds true for the SP algorithm:

$$\|\mathbf{x}_{T-T^\ell}\|_2 \leq \frac{1+\delta_{3K}}{1-\delta_{3K}} \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 + \frac{4}{1-\delta_{3K}} \|\mathbf{D}_{T_*}^* \mathbf{e}\|_2.$$

Proof: We will define the smear vector $\epsilon = \mathbf{x}_p - \mathbf{x}_{\tilde{T}^\ell}$, where \mathbf{x}_p is the outcome of the representation computation over \tilde{T}^ℓ , given by

$$\mathbf{x}_p = \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{y} = \mathbf{D}_{\tilde{T}^\ell}^\dagger (\mathbf{D}_T \mathbf{x}_T + \mathbf{e}), \quad (\text{C.1})$$

as defined in Algorithm 1. Expanding the first term in the last equality gives:

$$\begin{aligned} \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_T \mathbf{x}_T &= \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_{T \cap \tilde{T}^\ell} \mathbf{x}_{T \cap \tilde{T}^\ell} + \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_{T - \tilde{T}^\ell} \mathbf{x}_{T - \tilde{T}^\ell} \\ &= \mathbf{D}_{\tilde{T}^\ell}^\dagger [\mathbf{D}_{T \cap \tilde{T}^\ell}, \mathbf{D}_{\tilde{T}^\ell - T}] \begin{bmatrix} \mathbf{x}_{T \cap \tilde{T}^\ell} \\ \mathbf{0} \end{bmatrix} + \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_{T - \tilde{T}^\ell} \mathbf{x}_{T - \tilde{T}^\ell} \\ &= \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_{\tilde{T}^\ell} \mathbf{x}_{\tilde{T}^\ell} + \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_{T - \tilde{T}^\ell} \mathbf{x}_{T - \tilde{T}^\ell} \\ &= \mathbf{x}_{\tilde{T}^\ell} + \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_{T - \tilde{T}^\ell} \mathbf{x}_{T - \tilde{T}^\ell}. \end{aligned} \quad (\text{C.2})$$

The equalities hold based on the definition of $\mathbf{D}_{\tilde{T}^\ell}^\dagger$ and on the fact that \mathbf{x} is 0 outside of T . Using (C.2) we bound the smear energy from above, obtaining

$$\begin{aligned} \|\epsilon\|_2 &\leq \left\| \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{D}_T \mathbf{x}_T \right\|_2 + \left\| \mathbf{D}_{\tilde{T}^\ell}^\dagger \mathbf{e} \right\|_2 \\ &= \left\| (\mathbf{D}_{\tilde{T}^\ell}^* \mathbf{D}_{\tilde{T}^\ell})^{-1} \mathbf{D}_{\tilde{T}^\ell}^* \mathbf{D}_{T - \tilde{T}^\ell} \mathbf{x}_{T - \tilde{T}^\ell} \right\|_2 \\ &\quad + \left\| (\mathbf{D}_{\tilde{T}^\ell}^* \mathbf{D}_{\tilde{T}^\ell})^{-1} \mathbf{D}_{\tilde{T}^\ell}^* \mathbf{e} \right\|_2 \\ &\leq \frac{\delta_{3K}}{1-\delta_{3K}} \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 + \frac{1}{1-\delta_{3K}} \|\mathbf{D}_{\tilde{T}^\ell}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{C.3})$$

We now turn to bound $\|\epsilon\|_2$ from below. We denote the support of the K smallest coefficients in \mathbf{x}_p by $\Delta T \triangleq \tilde{T}^\ell - T^\ell$. Thus, for any set $T' \subset \tilde{T}^\ell$ of cardinality K , it holds that $\|(\mathbf{x}_p)_{\Delta T}\|_2 \leq \|(\mathbf{x}_p)_{T'}\|_2$. In particular, we shall choose T' such that $T' \cap T = \emptyset$, which necessarily exists because \tilde{T}^ℓ is of cardinality $2K$ and therefore there must be at K entries in this support that are outside T . Thus, using the relation $\epsilon = \mathbf{x}_p - \mathbf{x}_{\tilde{T}^\ell}$ we get

$$\begin{aligned} \|(\mathbf{x}_p)_{\Delta T}\|_2 &\leq \|(\mathbf{x}_p)_{T'}\|_2 = \|(\mathbf{x}_{\tilde{T}^\ell})_{T'} + \epsilon_{T'}\|_2 \\ &= \|\epsilon_{T'}\|_2 \leq \|\epsilon\|_2. \end{aligned} \quad (\text{C.4})$$

Because \mathbf{x} is supported on T we have that $\|\mathbf{x}_{\Delta T}\|_2 = \|\mathbf{x}_{\Delta T \cap T}\|_2$. An upper bound for this vector is reached by

$$\begin{aligned} \|\mathbf{x}_{\Delta T \cap T}\|_2 &= \|(\mathbf{x}_p)_{\Delta T \cap T} - \epsilon_{\Delta T \cap T}\|_2 \\ &\leq \|(\mathbf{x}_p)_{\Delta T \cap T}\|_2 + \|\epsilon_{\Delta T \cap T}\|_2 \\ &\leq \|(\mathbf{x}_p)_{\Delta T}\|_2 + \|\epsilon\|_2 \leq 2\|\epsilon\|_2, \end{aligned} \quad (\text{C.5})$$

where the last step uses (C.4). The vector \mathbf{x}_{T-T^ℓ} can be decomposed as $\mathbf{x}_{T-T^\ell} = [\mathbf{x}_{T \cap \Delta T}^*, \mathbf{x}_{T-\tilde{T}^\ell}^*]^*$. Using (C.3) and (C.5) we get

$$\begin{aligned} \|\mathbf{x}_{T-T^\ell}\|_2 &\leq \|\mathbf{x}_{T \cap \Delta T}\|_2 + \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 \leq 2\|\epsilon\|_2 + \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 \\ &\leq \left(1 + \frac{2\delta_{3K}}{1 - \delta_{3K}}\right) \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 + \frac{2}{1 - \delta_{3K}} \|\mathbf{D}_{\tilde{T}^\ell}^* \mathbf{e}\|_2 \\ &= \frac{1 + \delta_{3K}}{1 - \delta_{3K}} \|\mathbf{x}_{T-\tilde{T}^\ell}\|_2 + \frac{4}{1 - \delta_{3K}} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2, \end{aligned}$$

where the last step uses the property $\|\mathbf{D}_{\tilde{T}^\ell}^* \mathbf{e}\|_2 \leq 2\|\mathbf{D}_{T_e}^* \mathbf{e}\|_2$ taken from Section II, and this concludes the proof. \square

APPENDIX D

PROOF OF INEQUALITY (III.12)

Lemma D.1: The following inequality holds true for the CoSaMP algorithm:

$$\left\| \mathbf{x} - \hat{\mathbf{x}}_{\text{CoSaMP}}^\ell \right\|_2 \leq \frac{4\delta_{4K}}{(1 - \delta_{4K})^2} \left\| \mathbf{x} - \hat{\mathbf{x}}_{\text{CoSaMP}}^{\ell-1} \right\|_2 + \frac{14 - 6\delta_{4K}}{(1 - \delta_{4K})^2} \|\mathbf{D}_{T_e}^* \mathbf{e}\|_2.$$

Proof: We denote $\hat{\mathbf{x}}_{\text{CoSaMP}}^\ell$ as the solution of CoSaMP in the ℓ -th iteration: $\hat{\mathbf{x}}_{\text{CoSaMP}, (T^\ell)^c}^\ell = 0$ and $\hat{\mathbf{x}}_{\text{CoSaMP}, T^\ell}^\ell = (\mathbf{x}_p)_{T^\ell}$. We further define $\mathbf{r}^\ell \triangleq \mathbf{x} - \hat{\mathbf{x}}_{\text{CoSaMP}}^\ell$ and use the definition of $T_{\mathbf{e},p}$ (Section II).

Our proof is based on the proof of Theorem 4.1 and the Lemmas used with it [13].

Since we choose T_Δ to contain the biggest $2K$ elements in $\mathbf{D}^* \mathbf{y}_r^\ell$ and $|T^{\ell-1} \cup T| \leq 2K$ it holds true that $\|(\mathbf{D}^* \mathbf{y}_r^\ell)_{T^\ell \cup T}\|_2 \leq \|(\mathbf{D}^* \mathbf{y}_r^\ell)_{T_\Delta}\|_2$. Removing the common elements from both sides we get

$$\|(\mathbf{D}^* \mathbf{y}_r^\ell)_{(T^\ell \cup T) - T_\Delta}\|_2 \leq \|(\mathbf{D}^* \mathbf{y}_r^\ell)_{T_\Delta - (T^\ell \cup T)}\|_2. \quad (\text{D.1})$$

We proceed by bounding the rhs and lhs of (D.1), from above and from below respectively, using the triangle inequality. We use Propositions 2.1 and 2.2, the definition of $T_{\mathbf{e},2}$, and the fact that $\|\mathbf{r}^\ell\|_2 = \|\mathbf{r}_{T^\ell \cup T}^\ell\|_2$ (this holds true since the support of \mathbf{r}^ℓ is over $T \cup T^\ell$). For the rhs we obtain

$$\begin{aligned} \|(\mathbf{D}^* \mathbf{y}_r^\ell)_{T_\Delta - (T^\ell \cup T)}\|_2 &= \|\mathbf{D}_{T_\Delta - (T^\ell \cup T)}^* (\mathbf{D} \mathbf{r}^\ell + \mathbf{e})\|_2 \\ &\leq \|\mathbf{D}_{T_\Delta - (T^\ell \cup T)}^* \mathbf{D}_{T^\ell \cup T} \mathbf{r}_{T^\ell \cup T}^\ell\|_2 + \|\mathbf{D}_{T_\Delta - (T^\ell \cup T)}^* \mathbf{e}\|_2 \\ &\leq \delta_{4K} \|\mathbf{r}^\ell\|_2 + \|\mathbf{D}_{T_{\mathbf{e},2}}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{D.2})$$

and for the lhs:

$$\begin{aligned} \|(\mathbf{D}^* \mathbf{y}_r^\ell)_{(T^\ell \cup T) - T_\Delta}\|_2 &= \|\mathbf{D}_{(T^\ell \cup T) - T_\Delta}^* (\mathbf{D} \mathbf{r}^\ell + \mathbf{e})\|_2 \\ &\geq \|\mathbf{D}_{(T^\ell \cup T) - T_\Delta}^* \mathbf{D}_{(T^\ell \cup T) - T_\Delta} \mathbf{r}_{(T^\ell \cup T) - T_\Delta}^\ell\|_2 \\ &\quad - \|\mathbf{D}_{(T^\ell \cup T) - T_\Delta}^* \mathbf{D}_{T_\Delta} \mathbf{r}_{T_\Delta}^\ell\|_2 - \|\mathbf{D}_{(T^\ell \cup T) - T_\Delta}^* \mathbf{e}\|_2 \\ &\geq (1 - \delta_{2K}) \|\mathbf{r}_{(T^\ell \cup T) - T_\Delta}^\ell\|_2 - \delta_{4K} \|\mathbf{r}_{T_\Delta}^\ell\|_2 - \|\mathbf{D}_{T_{\mathbf{e},2}}^* \mathbf{e}\|_2. \end{aligned} \quad (\text{D.3})$$

Because \mathbf{r}^ℓ is supported over $T \cup T^\ell$, it holds true that $\|\mathbf{r}_{(T \cup T^\ell) - T_\Delta}^\ell\|_2 = \|\mathbf{r}_{T_\Delta^C}^\ell\|_2$. Combining (D.3) and (D.2) with (D.1), gives

$$\begin{aligned} \|\mathbf{r}_{T_\Delta^C}^\ell\|_2 &\leq \frac{2\delta_{4K} \|\mathbf{r}^\ell\|_2 + 2\|\mathbf{D}_{T_{\mathbf{e},2}}^* \mathbf{e}\|_2}{1 - \delta_{2K}} \\ &\leq \frac{2\delta_{4K} \|\mathbf{r}^\ell\|_2 + 4\|\mathbf{D}_{T_{\mathbf{e},2}}^* \mathbf{e}\|_2}{1 - \delta_{4K}}. \end{aligned} \quad (\text{D.4})$$

For brevity of notations, we denote hereafter \tilde{T}^ℓ as \tilde{T} . Using $\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{e} = \mathbf{D}_{\tilde{T}}\mathbf{x}_{\tilde{T}} + \mathbf{D}_{\tilde{T}^c}\mathbf{x}_{\tilde{T}^c} + \mathbf{e}$, we observe that

$$\begin{aligned}
\|\mathbf{x}_{\tilde{T}} - (\mathbf{x}_p)_{\tilde{T}}\|_2 &= \left\| \mathbf{x}_{\tilde{T}} - \mathbf{D}_{\tilde{T}}^\dagger (\mathbf{D}_{\tilde{T}}\mathbf{x}_{\tilde{T}} + \mathbf{D}_{\tilde{T}^c}\mathbf{x}_{\tilde{T}^c} + \mathbf{e}) \right\|_2 \\
&= \left\| \mathbf{D}_{\tilde{T}}^\dagger (\mathbf{D}_{\tilde{T}^c}\mathbf{x}_{\tilde{T}^c} + \mathbf{e}) \right\|_2 \\
&\leq \left\| (\mathbf{D}_{\tilde{T}}^*\mathbf{D}_{\tilde{T}})^{-1}\mathbf{D}_{\tilde{T}}^*\mathbf{D}_{\tilde{T}^c}\mathbf{x}_{\tilde{T}^c} \right\|_2 + \left\| (\mathbf{D}_{\tilde{T}}^*\mathbf{D}_{\tilde{T}})^{-1}\mathbf{D}_{\tilde{T}}^*\mathbf{e} \right\|_2 \\
&\leq \frac{1}{1 - \delta_{3K}} \left\| \mathbf{D}_{\tilde{T}}^*\mathbf{D}_{\tilde{T}^c}\mathbf{x}_{\tilde{T}^c} \right\|_2 + \frac{1}{1 - \delta_{3K}} \left\| \mathbf{D}_{T_e,3}^*\mathbf{e} \right\|_2 \\
&\leq \frac{\delta_{4K}}{1 - \delta_{4K}} \left\| \mathbf{x}_{\tilde{T}^c} \right\|_2 + \frac{3}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2,
\end{aligned} \tag{D.5}$$

where the last inequality holds true because of Proposition 2.4 and that $|\tilde{T}| = 3K$. Using the triangle inequality and the fact that \mathbf{x}_p is supported on \tilde{T} , we obtain

$$\|\mathbf{x} - \mathbf{x}_p\|_2 \leq \|\mathbf{x}_{\tilde{T}^c}\|_2 + \|\mathbf{x}_{\tilde{T}} - (\mathbf{x}_p)_{\tilde{T}}\|_2, \tag{D.6}$$

which leads to

$$\begin{aligned}
\|\mathbf{x} - \mathbf{x}_p\|_2 &\leq \left(1 + \frac{\delta_{4K}}{1 - \delta_{4K}} \right) \|\mathbf{x}_{\tilde{T}^c}\|_2 + \frac{3}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2 \\
&= \frac{1}{1 - \delta_{4K}} \|\mathbf{x}_{\tilde{T}^c}\|_2 + \frac{3}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2.
\end{aligned} \tag{D.7}$$

Having the above results we can obtain (III.12) by

$$\begin{aligned}
\left\| \mathbf{x} - \hat{\mathbf{x}}_{CoSaMP}^\ell \right\|_2 &\leq 2 \|\mathbf{x} - \mathbf{x}_p\|_2 \\
&\leq 2 \left(\frac{1}{1 - \delta_{4K}} \|\mathbf{x}_{\tilde{T}^c}\|_2 + \frac{3}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2 \right) \\
&\leq \frac{2}{1 - \delta_{4K}} \left\| \mathbf{r}_{T_\Delta^c}^{\ell-1} \right\|_2 + \frac{6}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2 \\
&\leq \frac{2}{1 - \delta_{4K}} \left(\frac{2\delta_{4K}}{1 - \delta_{4K}} \left\| \mathbf{r}^{\ell-1} \right\|_2 + \frac{4}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2 \right) + \frac{6}{1 - \delta_{4K}} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2 \\
&= \frac{4\delta_{4K}}{(1 - \delta_{4K})^2} \left\| \mathbf{r}^{\ell-1} \right\|_2 + \frac{14 - 6\delta_{4K}}{(1 - \delta_{4K})^2} \left\| \mathbf{D}_{T_e}^*\mathbf{e} \right\|_2,
\end{aligned} \tag{D.8}$$

where the inequalities are based on Lemma 4.5 from [13], (D.7), Lemma 4.3 from [13] and (D.4) respectively. \square

REFERENCES

- [1] D. L. Donoho, M. Elad, and V. N. Temlyakov, “Stable recovery of sparse overcomplete representations in the presence of noise,” *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 6–18, Jan. 2006.
- [2] D. L. Donoho and M. Elad, “On the stability of the basis pursuit in the presence of noise,” *Signal Process.*, vol. 86, no. 3, pp. 511–532, 2006.
- [3] J. Tropp, “Just relax: Convex programming methods for identifying sparse signals,” *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, Mar. 2006.
- [4] A. M. Bruckstein, D. L. Donoho, and M. Elad, “From sparse solutions of systems of equations to sparse modeling of signals and images,” *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.
- [5] B. Natarajan, “Sparse approximate solutions to linear systems,” *SIAM Journal on Computing*, vol. 24, pp. 227–234, 1995.
- [6] S. S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.
- [7] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc. B*, vol. 58, no. 1, pp. 267–288, 1996.
- [8] E. J. Candès and T. Tao, “The dantzig selector: Statistical estimation when p is much larger than n ,” *Annals Of Statistics*, vol. 35, p. 2313, 2007.
- [9] S. Chen, S. A. Billings, and W. Luo, “Orthogonal least squares methods and their application to non-linear system identification,” *International Journal of Control*, vol. 50, no. 5, pp. 1873–1896, 1989.
- [10] S. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Trans. Signal Process.*, vol. 41, pp. 3397–3415, 1993.
- [11] G. Davis, S. Mallat, and M. Avellaneda, “Adaptive greedy approximations,” *Journal of Constructive Approximation*, vol. 50, pp. 57–98, 1997.
- [12] D. Needell and R. Vershynin, “Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 310–316, april 2010.
- [13] D. Needell and J. Tropp, “Cosamp: Iterative signal recovery from incomplete and inaccurate samples,” *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.
- [14] W. Dai and O. Milenkovic, “Subspace pursuit for compressive sensing signal reconstruction,” *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.
- [15] T. Blumensath and M. E. Davies, “Iterative hard thresholding for compressed sensing,” *Applied and Computational Harmonic Analysis*, vol. 27, no. 3, pp. 265–274, 2009.
- [16] D. L. Donoho and X. Huo, “Uncertainty principles and ideal atomic decomposition,” *IEEE Trans. Inf. Theory*, vol. 47, no. 7, pp. 2845–2862, Nov 2001.
- [17] M. Elad and A. Bruckstein, “A generalized uncertainty principle and sparse representation in pairs of bases,” *IEEE Trans. Inf. Theory*, vol. 48, no. 9, pp. 2558–2567, Sep 2002.
- [18] D. L. Donoho and M. Elad, “Optimal sparse representation in general (nonorthogonal) dictionaries via l_1 minimization,” *Proceedings of the National Academy of Science*, vol. 100, pp. 2197–2202, Mar 2003.
- [19] E. J. Candès and T. Tao, “Near-optimal signal recovery from random projections: Universal encoding strategies?” *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5406–5425, dec. 2006.
- [20] E. Candès and T. Tao, “Decoding by linear programming,” *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, dec. 2005.
- [21] Z. Ben-Haim, Y. C. Eldar, and M. Elad, “Coherence-based performance guarantees for estimating a sparse vector under random noise,” *to appear in IEEE Trans. Signal Process.*, 2010.
- [22] M. Rudelson and R. Vershynin, “Sparse reconstruction by convex relaxation: Fourier and gaussian measurements,” in *Information Sciences and Systems, 2006 40th Annual Conference on*, 22–24 2006, pp. 207–212.
- [23] J. Tropp, “Just relax: convex programming methods for identifying sparse signals in noise,” *IEEE Trans. Inf. Theory*, vol. 52, no. 3, pp. 1030–1051, March 2006.

- [24] E. J. Candès, J. K. Romberg, and T. Tao, “Stable signal recovery from incomplete and inaccurate measurements,” *Communications on Pure and Applied Mathematics*, vol. 59, no. 8, pp. 1207–1223, 2006.
- [25] E. Candès, “The restricted isometry property and its implications for compressed sensing,” *Comptes Rendus Mathématique*, vol. 346, no. 9-10, pp. 589 – 592, 2008.
- [26] T. Cai, L. Wang, and G. Xu, “Shifting inequality and recovery of sparse signals,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1300 –1308, march 2010.
- [27] —, “New bounds for restricted isometry constants,” *To appear in IEEE Trans. Inf. Theory.*, 2010.
- [28] E. J. Candès, “Modern statistical estimation via oracle inequalities,” *Acta Numerica*, vol. 15, pp. 257–325, 2006.
- [29] P. J. Bickel, Y. Ritov, and A. B. Tsybakov, “Simultaneous analysis of lasso and dantzig selector,” *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.
- [30] T. Cai, L. Wang, and G. Xu, “Stable recovery of sparse signals and an oracle inequality,” *to appear in IEEE Transactions on Information Theory*, 2009. [Online]. Available: www-stat.wharton.upenn.edu/~tcai/paper/Stable-Recovery-MIP.pdf
- [31] Z. Ben-Haim and Y. C. Eldar, “The Cramèr-Rao bound for estimating a sparse parameter vector,” *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3384 –3389, june 2010.

